

المدونات اللغوية ودورها في معالجة النصوص العربية

دراسات ١٢

د.أيمن الدكروري

دراسات ١٢

المدونات اللغوية
ودورها في معالجة النصوص العربية

تأليف

د. أيمن الدكروري





المدونات اللغوية ودورها في معالجة النصوص العربية

الطبعة الأولى
١٤٣٩ هـ - ٢٠١٨ م

المملكة العربية السعودية - الرياض
ص.ب ١٢٥٠٠ الرياض ١١٤٧٣
هاتف:٠٠٩٦٦١١٢٥٨٧٢٦٨ - ٠٠٩٦٦١١٢٥٨١٠٨٢
البريد الإلكتروني: nashr@kaica.org.sa

مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية، ١٤٣٩هـ.
فهرسة مكتبة الملك فهد الوطنية أثناء النشر
الدكروري، أيمن
المدونات اللغوية ودورها في معالجة النصوص العربية/ أيمن الدكروري.- الرياض، ١٤٣٩ هـ
..ص؛ .. سم - (دراسات ؛ ١٢)
ردمك: ٩ -٠٩ -٨٢٢١ -٦٠٣ -٩٧٨
١ - اللغة العربية أ. العنوان ب. السلسلة
ديوي ٤١٠ ١٤٣٩/٣١٥٧
رقم الإيداع: ١٤٣٩/٣١٥٧
ردمك: ٩ -٠٩ -٨٢٢١ -٦٠٣ -٩٧٨

التصميم والإخراج
دار وجوه للنشر والتوزيع
Wojooh Publishing & Distribution House
www.wojoooh.com
المملكة العربية السعودية - الرياض
الهاتف:4562410 الفاكس:4561675
للتواصل والنشر:
info@wojoooh.com



الإهداء

إلى الغائبَين الحاضرَين

الدكتور السعيد محمد بدوي

و

الدكتور هانىء محيي الدين عطية

طيَّب الله ثراكما، وأسكنكما فسيح جناته!

فهرس الكتاب

المقدمة	١٧
الفصل الأول: المدونات اللغوية: الماهية والأهمية	٢٣
التمهيد	٢٥
المدونات اللغوية: المصطلح والمفهوم	٢٧
المدونات اللغويةوقواعد بيانات النصوص	٢٨
نظرةٌ تاريخيةٌ	٢٩
مميزات المدونات اللغوية	٣١
المواصفات العامة للمدونات اللغوية	٣٢
الأسئلة التي يمكن أن تجيب عنها المدونات اللغوية	٣٣
الأسئلة التي لا تجيب عنها المدونات اللغوية	٣٤
الخلاصة	٣٤

الفصل الثاني: المدونات اللغوية أداةً بحثيةً ومقاربةً منهجيةً	٣٥
التمهيد	٣٧
مجالات الإفادة من المدونات اللغوية	٣٨
الإفادة من المدونات اللغوية في علم اللغة	٣٨
الإفادة من المدونات اللغوية في تدريس وتعلم اللغات	٣٩
الإفادة من المدونات اللغوية في علم اللغة الاجتماعي Sociolinguistics	٤٢
الإفادة من المدونات اللغوية في صناعة المعاجم	٤٢
الإفادة من المدونات اللغوية في الترجمة	٤٦
الإفادة من المدونات اللغوية في دراسة التوجهات الفكرية (الأيديولوجيا)	٤٧
الإفادة من المدونات اللغوية في علم المعلومات	٤٨
الإفادة من المدونات اللغوية في صناعة المكانز	٥٠
الإفادة من المدونات اللغوية في المعلوماتية الحيوية Bioinformatics	٥٢
الإفادة من المدونات اللغوية في المعلوماتية الجنائية Forensic Informatics	٥٢
مراحل إنشاء المدونات اللغوية	٥٣
تحديد الهدف والأساس المنطقي rationale	٥٣
تحديد أنواع النصوص المناسبة	٥٤
تحديد مصادر النصوص	٥٤
الحصول على الموافقة القانونية	٥٤
جمع النصوص	٥٤
التحقق من النصوص التي تم جمعها	٥٦

حفظ النصوص في ملفات	٥٦
تشفير النصوص	٥٦
أنواع المدونات اللغوية	٥٧
مدوناتٌ لغويةٌ اختباريةٌ Test Corpora	٥٧
مدوناتٌ لغويةٌ بحثيةٌ Research Corpora	٥٧
مدونات لغويةٌ عامةٌ General Corpora	٥٧
مدونات لغويةٌ متخصصةٌ	٥٨
مدونات لغويةٌ تاريخيةٌ Historic أو تعاقبيةٌ Diachronic	٥٨
مدونات لغويةٌ راصدةٌ Monitor	٥٨
مدوناتٌ لغويةٌ تعليميةٌ Educational Corpora	٥٩
مدونات الدارسين (للغات) Learner Corpora	٥٩
مدونات لغويةٌ تربويةٌ أو تدريسيةٌ Pedagogical Corpora	٦٠
الاستخدام المنهجي للمدونات اللغوية	٦٠
الخلاصة	٦١
الفصل الثالث: معالجة البيانات اعتمادًا على المدونات اللغوية	٦٣
التمهيد	٦٥
عمليات معالجة البيانات على مستوى الإدخال	٦٦
ترميز المدونات اللغوية	٦٦
تحشية المدونات اللغوية	٦٧
أهمية تحشية المدونات اللغوية	٦٨

النقد الموجه لتحشية المدونات اللغوية	٦٩
طرق تحشية المدونات اللغوية	٧٠
أنواع تحشية المدونات اللغوية	٧٠
عمليات معالجة البيانات على مستوى الإخراج	٧٩
تكشيف الكلمات	٧٩
قوائم تردد الكلمات	٨٢
توليد الكلمات المفتاحية	٨٢
تحليل التجمعات العنقودية Cluster analysis	٨٤
الخلاصة	٨٧
الفصل الرابع: المدونات اللغوية: نماذج وبرمجياتٌ	٨٩
التمهيد	٩١
المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية King Abdulaziz City for Science and Technology (KACST) Arabic Corpus	٩٢
المدونة اللغوية التاريخية للجامعة الأردنية Historical Arabic Corpus	٩٩
المدونة اللغوية العربية الدولية لمكتبة الإسكندرية International Corpus of Arabic (ICA)	١٠١
مدونة عربي كوربَص Arabic Corpus	١٠٣
المدونة اللغوية لمتعلمي اللغة العربية Arabic Learner Corpus	١٠٤
المدونة العربية القرآنية The Quranic Arabic Corpus	١٠٦
مدونة قرآني Qurany	١٠٨
استقصاء المدونات اللغوية العربية Querying Arabic Corpora	١٠٩

سكتش إنجين Sketch Engine	١١٠
مخطط الكلمات Word Sketch	١١١
المكنز Thesaurus	١١٢
تكشيف النصوص Concordance	١١٣
قائمة الكلمات Word List	١١٣
استخراج المصطلحات Term Extraction	١١٤
تحليل المتتابعات اللفظية n-grams	١١٥
المدونة اللغوية للإنجليزية الأمريكية المعاصر The Corpus of Contemporary American English (COCA)	١١٦
المدونة اللغوية للأخبار على الشبكة العنكبوتية News on the Web (NOW Corpus)	١١٨
المدونة اللغوية لكتب جوجل Google Books Corpus	١١٨
برمجيات معالجة وتحليل المدونات اللغوية	١٢٠
برنامج أدوات وورد سميث WordSmith Tools	١٢٠
العمليات الفنية في برنامج أدوات وورد سميث	١٢١
برنامج أدوات معالجة المدونات اللغوية العربية The Arabic Corpus Processing Tools (ACPTs)	١٢٣
الخلاصة	١٢٦

الفصل الخامس: دراسات استخدام المدونات اللغوية	١٢٧
التمهيد	١٢٩
دراسات إنشاء وإتاحة المدونات اللغوية	١٣٠
دراسات الإفادة من المدونات اللغوية العربية في علم اللغة التطبيقي	١٣٦
في النحو والدلالة	١٣٦
في علم اللغة الاجتماعي	١٣٩
في صناعة المعاجم	١٤٢
في الترجمة	١٤٥
في دراسة التوجهات الفكرية (الأيديولوجيا)	١٤٦
دراسات استخدام المدونات اللغوية العربية في استرجاع المعلومات	١٤٨
دراسات الإفادة من المدونات اللغوية في صناعة المكانز	١٥١
دراسات استخدام المدونات اللغوية في المكتبات	١٥٣
الخلاصة	١٥٦
قائمة ببليوجرافية	١٥٧
أولًا: المراجع العربية	١٥٧
ثانيًا: المراجع الأجنبية	١٦٠
معجم المصطلحات	١٧٨
الكشاف	١٨٤

فهرس الأشكال

الشكل رقم (١) الصفحة الرئيسة للمدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية	٩٤
الشكل رقم (٢) صفحة البحث في المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية	٩٥
الشكل رقم (٣) صفحة البحث المخصص في المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية	٩٦
الشكل رقم (٤) صفحة توزيع التكرار في المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية	٩٧
الشكل رقم (٥) صفحة الكشاف السياقي في المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية	٩٨
الشكل رقم (٦) صفحة التصاحب اللفظي في المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية	٩٩

الشكل رقم (٧) الصفحة الرئيسة للمدونة اللغوية التاريخية للجامعة الأردنية	١٠٠
الشكل رقم (٨) الصفحة الرئيسة للمدونة اللغوية العربية الدولية	١٠٢
الشكل رقم (٩) الصفحة الرئيسة لمدونة عربي كوربَص	١٠٣
الشكل رقم (١٠) الصفحة الرئيسة للمدونة اللغوية لمتعلمي اللغة العربية	١٠٦
الشكل رقم (١١) التحشية الصرفية النحوية في المدونة العربية القرآنية	١٠٧
الشكل رقم (١٢) بنك الأشجار النحوي في المدونة العربية القرآنية	١٠٧
الشكل رقم (١٣) الأنطولوجيا الدلالية في المدونة العربية القرآنية	١٠٨
الشكل رقم (١٤) الصفحة الرئيسة للمدونة اللغوية «قرآني»	١٠٩
الشكل رقم (١٥) الصفحة الرئيسة لمدونة استقصاء المدونات اللغوية العربية	١١٠
الشكل رقم (١٦) واجهة الاستخدام الرئيسة لسكتش إنجين	١١١
الشكل رقم (١٧) إمكانات التحليل التي يقدمها مخطط الكلمات في سكتش إنجين لكلمة «team»	١١٢
الشكل رقم (١٨) صفحة المكنز المستخدم بسكتش إنجين لكلمة «argue»	١١٢
الشكل رقم (١٩) آلية عرض كشاف النصوص في سكتش إنجين	١١٣
الشكل رقم (٢٠) آلية عرض الكلمات وفقًا لتكرار ترددها بسكتش إنجين	١١٣
الشكل رقم (٢١) شاشة عرض المصطلحات والكلمات المفتاحية وفق تحليل سكتش إنجين	١١٤
الشكل رقم (٢٢) كيفية ضبط إعدادات تحليل المتتابعات اللفظية لاستخراج تسلسلات الكلمات والمتلازمات اللغوية	١١٥
الشكل رقم (٢٣) الصفحة الرئيسة لمدونة كوكا	١١٦
الشكل رقم (٢٤) صفحة البحث في مدونة كوكا	١١٧

الشكل رقم (٢٥) الصفحة الرئيسة للمدونة اللغوية «ناو»	١١٨
الشكل رقم (٢٦) الصفحة الرئيسة لمدونة كتب جوجل	١١٩
الشكل رقم (٢٧) واجهة الاستخدام الرئيسة لبرنامج أدوات وورد سميث	١٢٠
الشكل رقم (٢٨) طريقة الاختيار بين اختبار مربع كاي للدلالة الإحصائية، وطريقة احتمالات سجل الأداء في برنامج أدوات وورد سميث	١٢٢
الشكل رقم (٢٩) واجهات الاستخدام لبرنامج أدوات معالجة المدونات اللغوية العربية	١٢٤
الشكل رقم (٣٠) طريقة عرض التحليلات الإحصائية في برنامج غَوَّاص	١٢٥

السمات المعرفية

عدد هياكل الكلمات tokens	عدد الكلمات الفريدة types	نسبة هياكل الكلمات إلى الكلمات الفريدة type/token ratio
٣٦٥١٧	٨٣٠٤	٢٥,٠٠

الكلمات الوظيفية	**كلمات المحتوى**	**الكلمات المفتاحية**	
		المكونة من كلمتين	**المكونة من ٣ كلمات**
في	اللغوية	المدونات اللغوية	المدونات اللغوية العربية
من	المدونات	المدونة اللغوية	تحشية المدونات اللغوية
على	الكلمات	اللغة العربية	ترميز المدونات اللغوية
أو	النصوص	تكشيف النصوص	معالجة المدونات اللغوية
أن	المدونة	الكلمات المفتاحية	استخدام المدونات اللغوية
التي	اللغة		

الخريطة السحابية

اللغوية
المدونات
النصوص
الكلمات
المدونة

المقدمة

تحول الاقتصاد العالمي منذ ثمانينيات القرن العشرين من اقتصادٍ صناعيٍ إلى ما يُعرَف باقتصاد المعرفة؛ وذلك بصفته نتيجةً طبيعيةً للاستثمارات الكبيرة في مجالات البحث العلمي والتطبيق. الأمر الذي جعل المنتج الاستراتيجي الأبرز لأي أمةٍ هو المعرفة التي لا يمكن تناقلها إلا من خلال لغةٍ تحمل رسائلها عبر الأجيال والثقافات المختلفة.

ووفرت التكنولوجيا الحديثة، وعلى رأسها الإنترنت، في تسعينيات القرن الماضي الأرضَ الخصبة لنمو إنتاج اللغة بمعدلاتٍ مطَّردةٍ غير معهودةٍ، حتى أصبحت السوق هي العالم بأسره. الأمر الذي أدى إلى إنتاج عشرات الآلاف من الوثائق التي تحمل لغات العالم في كل دقيقةٍ. وقد صاحب ذلك كله إضافة قدرٍ من التعقيد على الصورة التي توجد عليها نصوص هذه اللغات، وجعلها متنوعةً بشكلٍ كبيرٍ. وهكذا أُضيفت أبعادٌ جديدةٌ وتبعاتٌ كثيرةٌ، واستُحدِثت أنماطٌ من مصادر معلوماتٍ لم نألفها من قبل. فأصبح الباحث تائهًا في خضم هذا الزحام وهذا التنوع.

وقد تنبه العالم إلى مخاطر اتساع الفجوة بين فيض النصوص والمعلومات هذا، وبين إمكانية تنظيمه والتحكم فيه. وهي فجوةٌ تعاني منها الكثير من لغات العالم، ولاسيما اللغة العربية. حتى باتت قضية تطويع التكنولوجيا الحديثة لخدمة اللغة العربية واحدةٌ من أهم القضايا الأكثر إلحاحًا، التي يتحتم تناولها ومعالجتها على النحو الذي من شأنه أن يرأب الصدع بين ما يتم إنتاجه من لغةٍ وبين إحكام السيطرة عليه.

ونحن أبناء اللغة العربية يتعين علينا المحافظة عليها بكافة السبل الممكنة؛ حتى تظل قويةً في مواجهة تحديات وصراعات العولمة. وحتى لا تُنسى أو تضيع ككثيرٍ من اللغات الأخرى. ولعل من أبرز السبل التي يمكن حذوها رصدُ واستكشافُ كل ما من شأنه أن يحفظ ويعالج ويتيح لغتنا، بما تحويه من نصوصٍ ومعلوماتٍ، وبما يلبي احتياجاتنا، على اختلاف مستوياتها وأنواعها، بالسرعة والسهولة المطلوبتين. غير أنه من الصعب جدًّا أن نُعَوِّل على الأدوات والطرق التقليدية التي يستحيل معها متابعة هذا الخضم الهائل من النصوص والمعلومات. ومن هنا دعت الحاجة إلى تطويع التكنولوجيا الحديثة للمعالجة والتطبيق مستفيدةً مما توفره إمكانات العصر من تقنياتٍ وبرمجياتٍ.

وتُعد المدونات اللغوية [Language or Textual] Corpora أحد السبل التي استفادت كثيرًا من التكنولوجيا الحديثة. وعلى الرغم من أن فكرة المدونات اللغوية قائمةٌ منذ آلاف السنين، إلا أن التكنولوجيا الحديثة أحيَّتها من جديدٍ. وأصبح هناك اهتمامٌ متزايدٌ في الآونة الأخيرة بمثل هذا المصدر من مصادر المعلومات.

فبعد أن كانت عمليات إنشاء وإتاحة المدونات اللغوية تشكل عبئًا وجهدًا كبيرين؛ لما تتطلبه من فرق عملٍ، إضافةً إلى الصعوبة البالغة في جمع البيانات الخام من آلاف النصوص- أصبحت هذه العمليات سهلة المنال نسبيًا بفضل التوجه السائد نحو رقمنة نصوص مصادر المعلومات الأولية ونشرها كي تصبح جزءًا من قواعد البيانات ومحتوى الشبكة العنكبوتية.

وسوف نحاول في هذا الكتاب، قدر الإمكان، أن نقدم لمدى إمكانية تطويع التكنولوجيا الحديثة لخدمة اللغة العربية باستخدام المدونات اللغوية. وذلك من خلال التعريف بهذه الأداة التقنية، والتحقق من أوجه الإفادة منها في مجالات المعرفة المختلفة. ويقع الكتاب في خمسة فصولٍ؛ ينصب **الفصل الأول** على ماهية وأهمية المدونات اللغوية. حيث يتم التحقق من هذا المصطلح ومفهومه، مع إلقاء نظرةٍ تاريخيةٍ على بدايات نشوء واستخدام المدونات اللغوية، ثم يتم تناول مميزات ومواصفات المدونات اللغوية، ويُختَتَم الفصل بالأسئلة التي يمكن أن تجيب عنها المدونات اللغوية، وما لم يمكن أن تجيب عنه.

فيما يتناول **الفصل الثاني** المدونات اللغوية بصفتها أداةً بحثيةً ومقاربةً منهجيةً. وفي هذا السياق يتم تسليط الضوء على مجالات الإفادة منها في كلٍ من: علم اللغة، وتدريس وتعلم اللغات، وعلم اللغة الاجتماعي، وصناعة المعاجم، والترجمة، ودراسة التوجهات الفكرية (الأيديولوجيا)، وعلم المعلومات، وصناعة المكانز، والمعلوماتية الجنائية. كما يتناول الفصل مراحل إنشاء المدونات اللغوية وأساليب تطويرها، وأنواعها المختلفة (مدونات لغويةٌ اختباريةٌ، ومدونات لغويةٌ بحثيةٌ، ومدونات لغويةٌ تعليميةٌ)، وكيفية استثمارها منهجيًا.

بينما يختص **الفصل الثالث** بمعالجة البيانات اعتمادًا على المدونات اللغوية. ويتطرق الحديث في هذا الموضوع إلى عمليات معالجة البيانات على مستويين؛ مستوى الإدخال، ومستوى الإخراج. ويتضمن المستوى الأول ترميزَ المدونات اللغوية، وتحشية المدونات اللغوية التي تؤدي دورًا كبيرًا في تحليل النصوص؛ ولذا سلط المؤلف الضوء على هذه العملية الفنية في ضوء تعريفها، وأهميتها، والنقد الموجه لها، وطرائقها، وأنواعها المختلفة التي تشمل: وسم أقسام الكلم، وتجريد الكلمات، والتحليل الإعرابي، والتحشية الدلالية، وتحشية المصاحبة المرجعية، والتحشية البراجماتية أو التداولية،

والتحشية الأسلوبية، ووسم الأخطاء، والتحشية الموجهة نحو المشكلات، والتحشية المتضَمنة، والتحشية القائمة بذاتها. بينما يشتمل المستوى الثاني تكشيف الكلمات، وتكشيف النصوص، وتكشيف الكلمات المفتاحية في السياق، والمشكلات الفنية في تكشيف المدونات اللغوية، والمشكلات اللغوية في تكشيف المدونات اللغوية، وقوائم تردد الكلمات، وتوليد الكلمات المفتاحية، وتحليل التجمعات العنقودية.

ويقدم **الفصل الرابع** نماذج فعليةً للمدونات اللغوية المعتمِدة على الشبكة العنكبوتية في إتاحتها وتقديم خدماتها. مع إبراز العربية منها قدر الإمكان. ومن بين هذه النماذج: المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية، والمدونة اللغوية التاريخية للجامعة الأردنية، والمدونة اللغوية العربية الدولية لمكتبة الإسكندرية، ومدونة عربي كوربَص، والمدونة اللغوية لمتعلمي اللغة العربية، والمدونة العربية القرآنية، ومدونة قرآني، واستقصاء المدونات اللغوية العربية، وسكتش إنجين، ومدونة كوكا، والمدونة اللغوية لكتب جوجل، والمدونة اللغوية للأخبار على الشبكة العنكبوتية (ناو). كما يقدم الفصل أشهر البرمجيات التي تعالج وتحلل المدونات اللغوية العربية، ومنها: برنامج أدوات وورد سميث WordSmith Tools، وبرنامج أدوات معالجة المدونات اللغوية العربية.

أما **الفصل الخامس** والأخير فيتعرض لدراسات استخدام المدونات اللغوية من خلال مراجعةٍ علميةٍ للإنتاج الفكري السابق حول الموضوع. مع التركيز على العربية فقط في معظم الأحيان. ويتناول هذا الإنتاج الفكري دراسات إنشاء وإتاحة المدونات اللغوية العربية، ودراسات الإفادة من المدونات اللغوية العربية في علم اللغة التطبيقي (في النحو والدلالة، وفي علم اللغة الاجتماعي، وفي صناعة المعاجم، وفي الترجمة، وفي تحليل التوجهات الفكرية)، ودراسات استخدام المدونات اللغوية العربية في صناعة المكانز، وفي استرجاع المعلومات، وفي المكتبات.

وبعد، فإننا نأمل أن ينتفع الدارسون، وعلماء اللغة، وعلماء المعلومات، وكافة الباحثين المهتمين بالمدونات اللغوية في شتى فروع المعرفة البشرية من هذا العمل المتواضع.

والله من وراء القصد،،،

الدكتور أيمن الدكروري

جدة

١٤٣٩ هـ/ ٢٠١٨ م

الفصل الأول

المدونات اللغوية: الماهية والأهمية

التمهيد
المدونات اللغوية: المصطلح والمفهوم
المدونات اللغوية وقواعد بيانات النصوص
نظرةٌ تاريخيةٌ
أهمية المدونات اللغوية
مميزات المدونات اللغوية
المواصفات العامة للمدونات اللغوية
الأسئلة التي يمكن أن تجيب عنها المدونات اللغوية
الأسئلة التي لا تجيب عنها المدونات اللغوية
الخلاصة

التمهيد

لا يمكن لنظم استرجاع النصوص، وما تشمله من أنماطٍ وأساليب مختلفةٍ للمعالجة الآلية للغة الطبيعية، أن تؤدي مهامها بكفاءةٍ، أو أن تتقدم وتفتح لنفسها مجالات تطبيقٍ متنوعةٍ، دون أن يتوافر لها وقودٌ خامٌ من مليارات البيانات يحرك هذه النظم، ويساعد الباحثين في الوصول إلى ما تحويه من مصطلحاتٍ وكلماتٍ، وفهم وتحليل ما بينها من علاقاتٍ، ليصبح هذا الوقود الخام مع الوقت مستودعًا للذاكرة، ومصدرًا متدفقًا للغة والمعلومات. وقد تجسد هذا الوقود الخام عمليًا فيما يُعرف بـ «المدونات اللغوية» Textual or Language Corpora التي تعتمد في عملها على قوة الحاسبات الآلية في الاختزان، وقدرات البرمجيات المختلفة في التكشيف، والتحليل، والاسترجاع.

ويُستخدم مصطلح «المدونات اللغوية» للدلالة على أي رصيدٍ ضخمٍ من النصوص، المكتوبة أو المنطوقة أو كلتيهما، التي يتم تجميعها بطريقةٍ عشوائيةٍ أو منظّمةٍ من مصادر النصوص المختلفة، ومن ثم يتم اختزانها في الحاسب الآلي لأغراض استرجاع المعلومات والرد على الاستفسارات وما شابه. إذن فالمدونات اللغوية تحوي نصوصًا تعكس الاستعمال الحقيقي أو الواقعي authentic للغة في شكلٍ مقروءٍ آليًا -machine readable تؤخذ عينةً ممثلةً لمجالٍ معينٍ، أو لأوعية معلوماتٍ بعينها، كالكتب، أو الدوريات العلمية، أو الصحف، أو المراجع... إلخ. وقد يَلحق بهذه النصوص ترميزٌ marking-up بإضافة حقول ميتاداتا، أو تحشيةٌ annotation، أو وسمٌ tagging (الزهيري، ٢٠٠٣، ص ص ٧٨-٧٩).

والمدونات اللغوية جمع مدونةٍ لغويةٍ؛ وهي في أبسط معانيها عبارةٌ عن مجموعةٍ مكونةٍ من وثيقتين أو أكثر (Khosrow-Pour, 2015, in Encyclopedia of Information Science and Technology, third edition, p. 4151).

والنصوص كما يعرفها معجم لونجمان لتدريس اللغة وعلم اللغة التطبيقي (Richard & Schmidt, 2002, p. 549) عبارةٌ عن لغةٍ منطوقةٍ أو مكتوبةٍ تتسم بالخصائص التالية:

١. تتكون في شكلها الطبيعي من عدةِ جملٍ تكون معًا تركيبًا أو وحدةً، كأن تكون مثلًا خطابًا، أو تقريرًا، أو مقالًا. وعلى الرغم من ذلك فإن الكلمة الواحدة تعتبر نصًا، كما هو الحال عند قراءة كلمة «خطر» بوصفها علامة تحذيرٍ.

٢. لها تركيبٌ مميزٌ وخصائص خطابيةٌ.

٣. تؤدي وظيفةً أو غرضًا تواصوليًا محددًا.

٤. تفهم غالبًا من خلال علاقتها بالسياق الذي تحدث فيه.

والنصوص في نظر دائرة المعارف الدولية لعلم المعلومات والمكتبات هي مجموعةٌ مترابطةٌ من الكلمات التي قد تكون نثرًا أو شعرًا، أو قائمةً من الكلمات، كما في الكشافات، أو عباراتٍ قصيرةً، وليس من الضروري أن تكون منتظمةً نحويًا، كما هو الحال في الشعارات الدعائية، وربما تكون أيضًا أرقامًا أو رموزًا شفهيةً (Feather & Sturges, 2003, p. 633).

ولمّا كانت اللغة أو النصوص، على وجه الخصوص، إما أن تكون مكتوبةً أو منطوقةً، باتت أيضًا المدونات اللغوية على هذا النحو. ومن أمثلة ذلك مدونة ويلنجتون للإنجليزية النيوزيلاندية المكتوبة The Wellington Corpus of Written New Zealand English (WWC)، ومدونة ويلنجتون للإنجليزية النيوزيلاندية المنطوقة The Wellington Corpus of Spoken New Zealand English (WSC) الصادرتان عن جامعة فيكتوريا ويلنجتون بنيوزيلاندا (Corpora at Victoria University of Wellington, 2013).

ويُذكر أن المدونات اللغوية، من حيث الوصول إليها، قد تكون متاحةٌ بالمجان على الشبكة العنكبوتية، مثل المدونة اللغوية لمتعلمي اللغة العربية Arabic Learner Corpus التي يقوم عليها الباحث عبد الله الفيفي، الباحث بجامعة الإمام محمد بن سعود

الإسلامية، المملكة العربية السعودية. وقد تكون المدونات اللغوية متاحةً بمقابلٍ مادي، مثل سكيتش إنجين Sketch Engine التي تتيح أيضًا الوصول إليها لفترةٍ تجريبيةٍ مدتها ثلاثون يومًا. كما تُتاح بعض المدونات اللغوية من خلال تسجيل البيانات للحصول على حساب استخدام مجانًا، مثل مدونة عربي كوربَص arabiCorpus الصادرة عن جامعة برجهام يَنج الأمريكية Brigham Young University arabiCorpus: the Arabic corpus for the rest of us).

وحريٌ بالذكر أن جميع الأمثلة المذكورة هنا للمدونات اللغوية قائمةٌ أو معتمدةٌ على الشبكة العنكبوتية web-based.

المدونات اللغوية: المصطلح والمفهوم

كلمة «corpus» كلمةٌ لاتينيةٌ تعني الجسد body، وجمعها corpora أو corpuses (Pearce, 2006, p. 45). ولكلمة «corpora» مقابلاتٌ عدةٌ في العربية، منها: المدونات اللغوية، والذخائر النصية (الربيعة، السلمان & إيريك، ٢٠١٤)، والذخائر اللغوية (ميدار: المشروع المتوسطي لتقنيات اللغات العربية المكتوبة والمنطوقة)، والمكتنزات النصية، والمتون اللغوية (حمادة، ٢٠١١)، وغيرها مما ذُكر في مصادر غير أكاديميةٍ، مثل: المُفَهرِسات، والمجموعات النصية، والمخزونات النصية... إلخ.

والمدونات اللغوية في حقيقة أمرها عبارةٌ عن رصيدٍ ضخمٍ من النصوص المكتوبة أو المنطوقة التي يتم اختزانها على الحاسب الآلي. وتعد الوثائق المنشورة في أحد فروع المعرفة مصدرًا رئيسًا للمدونة اللغوية، حيث تؤخذ الكلمات من العناوين، أو المستخلصات، أو الفقرات، أو النص الكامل؛ تمهيدًا لما سيتم عليها من تحليلٍ كمي ونوعي (Schneider, 2004, p. 27). وقد يتم تجهيز هذه النصوص ووضعها في شكلٍ معينٍ format يسمح بتعاملها مع أدوات البحث، وكذلك يسمح بعرض نتائج البحث فيها على نحوٍ يكفل قراءتها بسهولةٍ ويسرٍ من قِبل اختصاصيي المعلومات، واللغويين، والمدرسين، والطلاب، والمستفيدين. وفي هذه المرحلة يمكن طرح المدونات اللغوية كونها منتجًا يتم الاستفادة منه في مجالاتٍ عدةٍ، كلٌ وفق احتياجاته (Al-Sulaiti & Atwell, 2006, p. 5).

والمدونات اللغوية في حد ذاتها، حتى هذه المرحلة، لا تقدم شيئًا مطلقًا أكثر من كونها وعاءً أو وسيطًا إلكترونيًا لاختزان اللغة أو النصوص بصفتها بياناتٍ خامًا، إلى أن تأتي مرحلةٌ لاحقةٌ يتم فيها تشغيل بعض البرمجيات على هذه البيانات. وبمقدور هذه البرمجيات إعادة تنظيم وترتيب هذه البيانات. ومن ثم يتم عمل مجموعةٍ من التحليلات الإحصائية النظمية والدلالية عليها؛ بهدف فحص كل مصطلحٍ أو كلمةٍ مفتاحيةٍ من حيث سماتها أو تركيبتها المعرفية واللغوية، وما يرتبط بها من كلماتٍ تسبقها أو تلحقها. الأمر الذي يسهم بشكلٍ كبيرٍ في الصياغة المُثلى للمصطلحات، ورؤوس الموضوعات، والواصفات، وعبارات البحث، بشكلٍ مُحكَم يعكس الاستخدام الفعلي أو الواقعي authentic لها، وبعيدًا تمامًا عن البديهة أو الحدس. ويمكن بالتالي تقديم معلوماتٍ عن النصوص على النحو الذي يلبي احتياجات المستفيدين بدقةٍ وكفاءةٍ. وخلاصة القول أن المدونات اللغوية لا توفر معلومات عن النصوص، ولكن البرمجيات التي تعمل عليها هي التي تتولى ذلك (Hunston, 2002, p. 3).

المدونات اللغوية وقواعد بيانات النصوص

قد يَخلط البعض بين المدونات اللغوية وقواعد بيانات النصوص، أو نظم استرجاع النصوص، إلَّا أن الأخيرين لا يُدَلِلان على مفهوم المصطلح بشكلٍ دقيقٍ؛ إذ إنهما يوحيان بأن المقصود هو قاعدة بياناتٍ تُختزن بها بياناتٌ مكتوبةٌ written data ، قد تكون مُجَمَّعَةً بطريقةٍ عشوائيةٍ، ثم يتم إيعازها لاسترجاع النتائج بناءً على طلب بحثٍ من قِبل المستفيدين. وتكون عادةً هذه النتائج في صورة إشاراتٍ بعناوين مصادر المعلومات تُحيل الباحثين إلى النصوص الأصلية.

غير إن المدونات اللغوية يتم بناؤها وفقًا لأساسٍ منطقيٍ rationale محددٍ، كأن تكون المدونة اللغوية للإنتاج الفكري في علم اللغة فقط، أو في علم المعلومات فقط، أو في أي مجالٍ آخر، أو مجموعةٍ من المجالات المنتظمة والمصنفة داخل المدونات اللغوية. كما أنه قد يُختزن بها نصوصٌ مكتوبةٌ إضافةً إلى نصوص منطوقةٍ (مَحْكيَّةٍ). فضلًا عن أن البحث في المدونات اللغوية يتم بكلمة أو عبارة بحثٍ، فتُسترجع النتائج التي تشتمل على هذه الكلمة أو العبارة في سياقاتها الطبيعية (تكشيف النصوص)، دون ضرورة إحالة المستفيدين إلى النصوص الكاملة لمصادر المعلومات في الغالب. ونضيف إلى ذلك أن

المدونات اللغوية تتيح عادةً إمكانات استرجاع غير متوافرةٍ في قواعد بيانات النصوص أو نظم استرجاع النصوص؛ مثل قوائم تردد ألكلمات، واشتقاق (أو توليد) الكلمات المفتاحية، وعرض الأشكال المختلفة لكلمة البحث، وعرض الكلمات التي تتلازم أو تتصاحب عادةً مع كلمة أو عبارة البحث. الأمر الذي يفيد كثيرًا في إجراء وتنفيذ الكثير من الدراسات والمشروعات المختلفة، كما سيتضح تباعًا (المالكي، ٢٠٠٩، ص ص ٥-٦).

ومن الممكن النظر إلى المدونات اللغوية كونها مُكَوِّناً رئيسًا تعتمد عليه نظم استرجاع النصوص، أو نظم استرجاع المعلومات بوجهٍ عامٍ؛ سواءٌ في مراحل إنشائها، أو مراحل اختبارها قبل طرحها للمستفيدين. وسيتم توضيح هذا الأمر عند الحديث عن الإفادة من المدونات اللغوية في علم المعلومات، ومعالجة اللغة الطبيعة للنصوص.

نظرةٌ تاريخيةٌ

ارتبطت فكرة تجميع النصوص بتكشيف النصوص الدينية للإنجيل في القرنين السابع والثامن للميلاد، ومن قبلها النصوص الدينية للتوراة (Borko & Bernier, 1978). وكذلك النصوص التاريخية، مثل لفافات البحر الميت التي تُعد من أهم مصادر تاريخ فلسطين (قاسم، ٢٠٠٠، ص ٢٢).

وربما تعود أول محاولةٍ لإنشاء مدونةٍ لغويةٍ متخصصةٍ اعتمادًا على الحاسب الآلي إلى الأب روبرتو بوسا Roberto Busa حينما قام بتجميع أعمال القديس توما الأكويني Thomas Aquinas في عام ١٩٤٩ وتكشيف نصوصها باستخدام أجهزة البطاقات المثقبة (Winter, 1999).

بينما تعود أول محاولةٍ لإنشاء مدونةٍ لغويةٍ عامةٍ باستخدام الحاسب الآلي إلى الباحثين كيوسيرا Kucera وفرانسيز Francis في عام ١٩٦٠ حينما قاما بإنشاء مدونةٍ لغويةٍ لأكثر من مليون كلمةٍ (هيكل كلمة[1] token) للغة الإنجليزية الأمريكية. وقد عُرفت هذه المدونة اللغوية باسم مدونة براون Brown corpus. وقد اقتصرت بداية هذه المدونة اللغوية على تكشيف وحساب تردد الكلمات فقط، إلا أن القائمين عليها قاموا مؤخرًا بِوَسْم أقسام كلماتها Part-Of-Speech tagging (Francis & Kucera, 1979). ثم توالى بعد ذلك إنشاء الكثير من المدونات اللغوية الأخرى، التي منها، على سبيل المثال

١- هي سلسلةٌ من الحروف، أو التمثيلات، أو كلتيهما، يسبقها فراغٌ ويتبعها فراغٌ.

لا الحصر، المدونة اللغوية الوطنية البريطانية British National Corpus (BNC)، والمدونة اللغوية الوطنية الأمريكية American National Corpus (ANC)، وغيرهما...

كما أُنشئت أيضًا مدوناتٌ لغويةٌ للأعمال الأدبية البارزة ومعالجتها، كما هو الحال في كشاف نصوص أشعار ماثيو أرنولد A concordance to the poems of Matthew Arnold الذي أعده باريش S. M. Parrish، ونشرته جامعة كورنيل Cornell عام ١٩٥٩ (عبدالهادي و زايد، ٢٠٠٠، ص ص ٨٤-٨٥). وكشاف نصوص أعمال وليم شكسبير Complete and Systematic Concordance to the Works of Shakespeare في عام ١٩٧٠ (قاسم، ٢٠٠٠، ص ٢٣).

وعلى الصعيد العربي، ربما تعود فكرة تجميع النصوص إلى النحاة الأوائل الذين استنبطوا ووضعوا قواعد اللغة العربية ومعاجمها اعتمادًا على ما كانوا يجمعونه من نصوصٍ عربيةٍ، أو ربما ترجع الفكرة إلى تجميع الأعمال الكاملة للمؤلفين (عمر، ١٩٨٨، ص ٥٦).

كما يُرجِع البعض فكرة تجميع النصوص إلى الجهود الرامية لتوفير مفاتيح أو أطراف[1] الوصول للأحاديث النبوية في مظانها في القرن الأول الهجري، على يد ابن سيرين - ١١٠ هـ. وظهرت بعد ذلك مجموعةٌ أخرى من أرصدة النصوص لتكشيف دواوين الشعر، وصناعة المعاجم اللغوية، مثل «كتاب العين» للخليل بن أحمد - ١٧٥هـ- وصناعة معاجم القرآن الكريم. واعتمدت كلها بشكلٍ مباشرٍ على أسس وآليات تكشيف النصوص (عرفات، ٢٠٠٩، ص ص ٥١-٦٨).

وجديرٌ بالذكر أن المسلمين الأوائل لم يكونوا في حاجةٍ إلى تجميع القرآن الكريم أو الأحاديث النبوية الشريفة؛ فقد اعتمدوا في البداية على الحُفاظ والرواة. إلا أن الأمر لم يدم كثيرًا، لاسيما بعد انتشار المصاحف المطبوعة وقلة حفظة القرآن والحديث، واستحداث فروعٍ جديدةٍ من المعرفة، فأصبحت الحاجة ملحةً إلى تجميع نصوص القرآن الكريم والأحاديث النبوية.

وقد كان لتجميع نصوص القرآن الكريم وتكشيفها منهجان؛ أحدهما إسلاميٌّ، والآخر أوربيٌّ استشراقيٌّ. وربما تعود أول محاولةٍ في هذا الصدد إلى الورادي حافظ

١- حيث كان يُدوَن طرف من الحديث للتذكير به، أو للدلالة أو للتوصيل إلى بقيته في كتابٍ من كتب الحديث أو عدة كتبٍ منها.

إبراهيم، كما أشار إليه إبراهيم الإبياري في الموسوعة الإسلامية، الذي وضع فهرسه المعروف باسم «ترتيب زيبا» الذي رُتبت فيه آيات القرآن الكريم على نمطٍ يخالف النهج المعجمي؛ حيث اعتمد على أوائل الآيات. إلى أن جاء حافظ إبراهيم بن مصطفى وحاول إعادة النظر في «ترتيب زيبا» وتيسير الانتفاع به؛ فوضع كتابه «تسهيل الترتيب».

أما المنهج الثاني الذي سلكه العرب والمسلمون بشأن تجميع وتكشيف النصوص، فقد أخذوه عن المستشرقين. حيث استفاد هؤلاء الغربيون مما يوفره تكشيف النصوص من إمكاناتٍ؛ فقام المستشرق الألماني جوستاف فلوجل بتكشيف نصوص القرآن الكريم في كتابه الموسوم «نجوم الفرقان في أطراف القرآن» الذي رتب فيه ألفاظ القرآن ترتيبًا ألفبائيًا على حروف المعجم. ثم وضع علمي زاده فيض الله الحسيني كتابه «فتح الرحمن لطالب آيات القرآن» الذي رتب فيه الكلمات ترتيبًا معجميًا، ووضع الكلمات تحت رؤوس موادها، ثم وضع رمزًا للسور، وترك الكلمات التي يكثر ترددها.

ثم جاء محمد فؤاد عبد الباقي وسار على نهج فلوجل، وصحح الأخطاء التي وقع فيها، واستفاد من عمل زاده فتجنب استخدام الرموز المعقدة. حيث رُتبت ألفاظ القرآن الكريم الواردة بالمصحف العثماني ترتيبًا هجائيًا وفقًا للمواد اللغوية، وتحت كل مادةٍ الألفاظُ المشتقة منها، ثم رُبطت الألفاظ بالآيات التي وردت بها ثم أسماء السور. وقد خرج هذا العمل بعنوان «المعجم المفهرس لألفاظ القرآن الكريم» في عام ١٩٤٥.

ومن الجهود الحديثة للاستفادة من المدونات اللغوية «المعجم المفهرس لألفاظ الحديث النبوي عن الكتب الستة وعن مسند الدارمي وموطأ مالك ومسند أحمد بن حنبل» الذي أشرف على إعداده المستشرق الهولندي فنسنك في عام ١٩٣٦ (قاسم، ٢٠٠٠، ص ص ٢٣-٢٤).

مميزات المدونات اللغوية

تتسم المدونات اللغوية بمميزاتٍ عدةٍ، لعل من أهمها:

١. الاعتماد بالأساس على نصوصٍ واقعيةٍ أو فعليةٍ authentic، وبالتالي فإنها تتيح إمكانية التحقق من صحة النتائج المبنية على الحدس أو التخمين في ضوء الاستعمال الواقعي للنصوص. وبمعنىً آخر، فإن المدونات اللغوية تجعل الباحثين في موقفٍ يقيني من نتائج البحث؛ لأن نتائجها المسترجعة تعتمد بالأساس على نصوصٍ كما وردت بالفعل

في سياقاتها الطبيعية، كما يستخدمها أبناء اللغة أو أهل التخصص، وبذلك فهي بعيدةٌ كلَ البُعد عن الظن أو البديهة (Bowker & Pearson, 2002, pp. 9-19).

٢. كبر حجم النصوص القابلة للبحث والتحليل. ويُحسب الحجم هنا بعدد الكلمات، أو بالأحرى هياكل الكلمات tokens، التي تحويها المدونات اللغوية. ويُقدر الحجم غالبًا بملايين الكلمات للمدونات اللغوية العامة. ولكن ذلك قد لا ينطبق بالضرورة على المدونات اللغوية المتخصصة (Bowker & Pearson, 2002, pp. 45-48).

٣. التنوع المبني على أسسٍ علميةٍ لنصوص المدونات اللغوية لتمثل استخدامات النصوص المختلفة؛ وذلك بمراعاة التمثيل الجغرافي والتاريخي والنوعي (الأسلوبي مثلًا) للنصوص واستعمالاتها المختلفة. غير إن ذلك يعتمد بالطبع على التصميم والاختيار الجيدين من قِبل معدي المدونات اللغوية (Bowker & Pearson, 2002, pp. 10-11).

المواصفات العامة للمدونات اللغوية

من الأمور التي يتعين وضعها في الحسبان عند إنشاء المدونات اللغوية وتقويمها ما يلي:

١. الحجم الكبير big size: يعتمد حجم المدونة اللغوية بشكلٍ أساسٍ على نوع الأسئلة التي تُسهم المدونة اللغوية في الإجابة عنها، والأهداف التي تسعى التحليلات التي تتم على المدونة اللغوية إلى تحقيقها. وبوجهٍ عام، فإنه يُفضل الحجم الكبير للمدونات اللغوية؛ وخصوصًا إذا علمنا أن الكثير من الكلمات والمتلازمات اللغوية تتردد بتكراراتٍ منخفضةٍ (Al-Thubaity, 2015, p. 727)، ولما يمكن أن يسهم به الحجم الكبير للمدونات في اضطلاع برامج معالجة النصوص بمهامها وتنقيح نتائجها بسهولةٍ ويسرٍ. وعلى الرغم من ذلك فإنه من الممكن أن تكون المدونات اللغوية صغيرة الحجم أفضل، ولاسيما إذا ما أُريد التحقق من الكلمات كثيرة التردد. فتكون حينها المدونات اللغوية صغيرة الحجم أكثر ملاءمة من تلك المفرطة في كبر حجمها. مع ضرورة الوضع في الحسبان كذلك أن بعض برامج معالجة النصوص تُوقِف عدد أسطر الكشاف المسترجعة عند حدودٍ معينةٍ، بحيث عندما تصل إلى حدٍ معينٍ، فإنها تتوقف عن الاستمرار في البحث داخل المدونة اللغوية.

وعلاوةً على ذلك، فإن هناك اعتباراتٍ عمليةً أخرى. فعلى سبيل المثال، إذا أُريد استعمال مدونةٍ لغويةٍ منطوقةٍ ذات تفصيلاتٍ عاليةٍ من الدقة، فإنه من الأجدى أن

يتم التعامل مع آلافٍ من الكلمات، وليس ملايين من الكلمات. بينما يعتمد الموقف مع المدونات اللغوية المكتوبة على الحصول على تصاريح مسبقةٍ حتى لا تُنتهك حقوق الملكية. (Evans, 2017).

٢. الشمولية والتمثيل comprehensiveness and representativeness: تعني مدى تغطية العينة المنتقاة للتغير الواقع في مجتمع النصوص التي يتم تحليلها. ويتضمن ذلك نوع، وعدد، وطول، واختيار عينات النصوص التي يتم تغطيتها. ويُسبَق كل ذلك بقراراتٍ منهجيةٍ لتحديد مجتمع النصوص وأخذ العينة sampling (Biber, 1994).

٣. التوازن balance: أي أن يكون هناك توازنٌ بين أنواع أو فئات النصوص والتخصصات وغير ذلك مما يشمله معيار التمثيل، فلا يطغى مؤلفٌ، أو منطقةٌ جغرافيةٌ، أو فترةٌ زمنيةٌ، أو نطاقٌ... على غيره. الأمر الذي يضمن الموضوعية والبُعد عن التحيز، وبالتالي إمكانية الخروج بمؤشراتٍ ونتائج أكثر دقةً وتمثيلًا لمجتمع الدراسة (McEnery, 2006).

الأسئلة التي يمكن أن تجيب عنها المدونات اللغوية

يمكن للمدونات اللغوية أن تجيب عن الأسئلة التالية:

١. ما أكثر الكلمات أو العبارات ترددًا؟

٢. ما أوجه الاختلاف بين النصوص المكتوبة والنصوص المنطوقة؟

٣. ما الأفعال، أو الأسماء، أو الحروف التي يستخدمها أهل اللغة أو أهل التخصص أكثر من غيرها؟

٤. ما حروف الجر (أو الأفعال، أو الأسماء) التي تسبق أو تلي كلمةً بعينها؟

٥. كيف يستخدم أهل اللغة أو أهل التخصص كلمةً أو مصطلحًا معينًا؟

٦. كم مرة تُستخدم فيها التعبيرات الاصطلاحية بين أهل اللغة أو أهل تخصصٍ ما؟

وغيرها الكثير من الأسئلة التي يمكن طرحها وفق الهدف الذي يحدده المستفيدون أنفسهم من المدونات اللغوية (McCarthy, 2004, pp. 1-2).

الأسئلة التي لا تجيب عنها المدونات اللغوية

إذا كان من المهم فهم ما يمكن أن تقدمه المدونات اللغوية من خدماتٍ جليلةٍ، فإنه من المهم أيضًا أن نعرف ما ليس بمقدورها أن توفره أو تجيب عنه. فالمدونات اللغوية لا يمكنها أن تجيب عن الأسئلة التالية:

١. ما البراهين أو الأدلة السلبية حول استعمال كلمةٍ، أو مصطلحٍ، أو عبارةٍ معينةٍ؟
فالمدونات اللغوية لا تقدم ما هو ممكنًا أو مناسبًا، أو غير ممكنٍ أو غير مناسبٍ من النصوص؛ بل إنها تقدم فقط ما هو متاحًا أو غير متاحٍ في المدونة اللغوية. ولذا نجد أن البعض قد يخطىء في اعتقاده بأن المدونات اللغوية لا تقدم كافة الأساليب والطرق التي تُعبر عن فكرةٍ معينةٍ، وبالتالي فإنهم لا يثقون في البراهين والأدلة التي تقدمها المدونات اللغوية. غير إن هؤلاء ينبغي أن يضعوا في الحسبان أنه إذا لم يكن أسلوب الفكرة ممثَلًا فيما يتم تقديمه، فلربما يرجع ذلك إذن إلى أن هذا الأسلوب غير شائع الاستخدام في الموضوع أو النوع الأدبي الذي تغطيه المدونة اللغوية المعنية.

٢. لماذا...؟
فالمدونات اللغوية لا يمكنها تفسير: «لماذا» هذه الظاهرة اللغوية، أو هذا المصطلح، أو غيرهما؟ فالإجابة عن «لماذا» لا تتم إلا عن طريق أهل اللغة أو أهل التخصص أنفسهم باستخدام حدسهم أو بديهتم.

٣. ما كافة الاستخدامات الممكنة لمصطلحٍ أو كلمةٍ أو عبارةٍ في اللغة على إطلاقها؟
فالمدونات اللغوية لا يمكنها توفير كافة الاستخدامات الممكنة للغةٍ في آنٍ واحدٍ. إذ إن المدونات اللغوية، مهما بلغ حجمها، فإنه لا يمكنها أبدًا تغطية كافة السياقات الممكنة التي تُستخدم فيها اللغة، بل إنها تظل دائمًا معنيةً بنوعٍ أو موضوعٍ محددٍ، وليس كافة الأنواع والموضوعات (Bennett, 2010, pp. 1-2).

الخلاصة

انصب هذا الفصل على موضوع المدونات اللغوية من حيث الماهية والأهمية. وفي ضوء ذلك تم سرد الأصل الإنجليزي للكلمة، ومقابلاتها المختلفة في اللغة العربية، ومفهومها، مع إلقاء نظرةٍ تاريخيةٍ على بدايات نشوء واستخدام وتطور المدونات اللغوية، والتعرف على مميزات ومواصفات المدونات اللغوية، وما يمكن للمدونات اللغوية أن تجيب عنه، وما تعجز أن تقدمه.

الفصل الثاني

المدونات اللغوية أداةً بحثيةً ومقاربةً منهجيةً

التمهيد
مجالات الإفادة من المدونات اللغوية
الإفادة من المدونات اللغوية في علم اللغة
الإفادة من المدونات اللغوية في تدريس وتعلم اللغات
الإفادة من المدونات اللغوية في علم اللغة الاجتماعي
الإفادة من المدونات اللغوية في صناعة المعاجم
الإفادة من المدونات اللغوية في الترجمة
الإفادة من المدونات اللغوية في دراسة التوجهات الفكرية (الأيديولوجيا)
الإفادة من المدونات اللغوية في علم المعلومات
الإفادة من المدونات اللغوية في صناعة المكانز
الإفادة من المدونات اللغوية في المعلوماتية الحيوية
الإفادة من المدونات اللغوية في المعلوماتية الجنائية
مراحل إنشاء المدونات اللغوية
أنواع المدونات اللغوية
الاستخدام المنهجي للمدونات اللغوية
الخلاصة

التمهيد

بعدما تم التحقق من ماهية وأهمية المدونات اللغوية، نحاول في هذا الفصل من الكتاب أن نسلط الضوء على تلك الأداة التي يسير استخدامها واستثمارها بخطىً متسارعةٍ في فروع عدةٍ من فروع المعرفة البشرية. وفي هذا السياق سيتم تناول المدونات اللغوية، كونها أداةً في إجراء البحوث العلمية، ومقاربةً منهجيةً في إعداد الدراسات في مجالات المعرفة المختلفة، إضافةً إلى مراحل إنشائها، وأنواعها، واستخدامها منهجيًا.

مجالات الإفادة من المدونات اللغوية

تُعد المدونات اللغوية مصدرًا حيويًا للبيانات التي يتم تحليلها إذا ما أُريد إجراء بحثٍ أو تقييمٍ في عددٍ من التخصصات. ولذا سيتم الحديث هنا عن استخدامات المدونات اللغوية في مجالات المعرفة التي استثمرت بالفعل هذه الأداة في النهوض بتطبيقاتها العملية.

الإفادة من المدونات اللغوية في علم اللغة

من الملاحظ أن معظم دراسات علم اللغة العام دراساتٌ وصفيةٌ أكثر منها إرشادية. فضلًا عن أنها لا تستخدم الطرق التجريبية في منهجياتها، بل تعمد كثيرًا إلى البديهة أو الحدس في التعامل مع اللغة. علاوةً على أن الدراسات الوصفية للغة قد تغفل الكثير من التنوع في استخدام اللغة ذاتها. الأمر الذي يجعل مثل هذه الدراسات تفتقد إلى القدر المأمول من الموضوعية. ولذلك يؤكد العديد من علماء اللغة على أهمية الاعتماد على المدونات اللغوية في إجرء بحوث اللغة بغية تلافي كل هذه المعضلات، والنأي بنتائج الدراسات عن أي شكلٍ من أشكال التحيز أو الآراء والأمثلة التطبيقية المتأثرة بالاعتبارات الشخصية (Hunston, 2002, pp. 13-14; Alansary, Nagi & Adly, 2008, p. 19).

وربما يكون أكثر المجالات إفادةً من المدونات اللغوية بوجهٍ عامٍ هو علم اللغة التطبيقي الذي شرع في استثمار هذه الأداة من أجل النهوض بالكثير من الدراسات اللغوية على مختلف مستوياتها؛ الصوتي، والصرفي، والنحوي، والدلالي، والخطابي، والمعجمي. إذ يمْكن الاعتماد على المدونات اللغوية في الحصول على معلوماتٍ تعكس الاستخدام أو الواقع الفعلي authentic للظواهر اللغوية. وفي الوقت نفسه تُستخدم هذه المعلومات كونها أساسًا في تجهيز وإعداد كتب قواعد اللغة. ولذا شهدت العقود الثلاثة الأخيرة زيادةً ملحوظةً في الإفادة من تطبيقات المدونات اللغوية في هذا المجال. (Hunston, 2002, pp. 13-14; Leech, 1997, p. 9; Biber, et al, 1998, McEnery, Wilson, 2001; Alansary, Nagi & Adly, 2008, p. 19).

فالمدونات اللغوية بالنسبة لعلم اللغة التطبيقي بمثابة أداةٍ، مثلها مثل التلسكوب في أهميته ودوره في علوم الفضاء. وهي في الوقت ذاته ليست كالمجهر أو الميكروسكوب؛ ولذا لا ينبغي أن تُنتقد بسبب عجزها عن القيام ببعض المهام. وبمعنىً آخر فإن

المدونات اللغوية من شأنها رصد وتقريب الظواهر اللغوية المختلفة من خلال تجميع واسترجاع كافة سياقاتها. غير أنها لا يمكنها تكبير الظواهر اللغوية الدقيقة التي يستلزم تفسير وتحليلها؛ فالعنصر البشري وحده هو القادر على أداء مثل هذه المهام (Stubbs, 1996, p. 231; Stubbs, 1999; Hunston, 2002, p. 20).

الإفادة من المدونات اللغوية في تدريس وتعلم اللغات

أضحت المدونات اللغوية أداةً رئيسةً في يد الكثير من معلمي ودارسي اللغات. غير إن اللغة العربية ما تزال مفتقدةً لاستثمار مثل هذه الأداة كما ينبغي أن يكون. حيث من الممكن أن تُفيد المدونات اللغوية كثيرًا في مجال تدريس وتعلم اللغات من جانب الباحثين من ناحيةٍ، ومن جانب الطلاب من ناحيةٍ أخرى، ومن جانب المعلمين من ناحيةٍ ثالثةٍ.

فيما يخص الجانب الأول فإنه يمكن الإفادة من المدونات اللغوية في ضوء مساراتٍ عدةٍ؛ لعل أهمها التحقق من:

١. التركيب اللغوي والمعنى pattern and meaning:

يُعد التركيب اللغوي والمعنى واحدًا من الملاحظات المحورية التي يمكن استخراجها من نصوص المدونات اللغوية. فقد يكون لكلمةٍ واحدةٍ أكثر من معنىً أو دلالةٍ، وكلٌ منها يرتبط بتركيبٍ لغويٍ يميزه (Sinclair, 1991). فعلى سبيل المثال عندما يُستخدم الظرف «قبل» ليدل على ترتيب الأحداث في شكل علاقة السبب والنتيجة، فإنه يُسبق عادةً بفعلٍ مضارعٍ أو فعلٍ ماضٍ، ويُتبع عادةً بمصدرٍ صريحٍ أو مصدرٍ مؤولٍ، ويظل مستمسكًا بمعناه المباشر، كما في: تحطمت في الجو قبل هبوطها. غير إن هذا الاسم نفسه عندما يُستخدم للدلالة على فترةٍ زمنيةٍ فإنه يُسبق بفعلٍ ماضٍ، ويُتبع باسمٍ (عدد)، ويأخذ معنىً آخر (منذ)، كما في: أكد قبل ساعاتٍ.

وبناءً عليه، فإنه لا يمكن بحالٍ القول بأن المعنى الواحد يرتبط بكلمةٍ واحدةٍ، ولكن التركيب اللغوي أو الأسلوب العباري للكلمات phraseology ككل هو القادر على تحديد المعنى.

وقد يحدث العكس عند التحليل؛ أي يتم تحديد المعنى المشترك أولًا، ثم يتم التحقق من التركيب اللغوي الذي يعطي هذا المعنى (Eddakrouri, 2016, p. 119).

٢. التعبيرات الاصطلاحية idiomatic expressions:

يُنظر دائمًا إلى اللغة على أنها مجموعةٌ من الكلمات التي تسهم في تقديم معنىً معينٍ. ولا يمكن التحقق من هذا المعنى إذا ما تم النظر إلى هذه الكلمات كونها وحداتٍ معجميةً مفردةً أو منفصلةً. فضلًا عن أنه في بعض الحالات، أو ربما كثيرٍ منها، يُستخدم أسلوبٌ عباريٌ معينٌ في تقديم معنىً مغايرٍ لما هو متوقعٍ. وهنا تلعب المدونات اللغوية دورًا فعالًا في حصر واسترجاع كافة السياقات التي تضم مثل هذا الأسلوب العباري للكلمات.

فعلى سبيل المثال يُستخدم الظرف «قبل» بمعناه المباشر في معظم الحالات، إلا أنه عندما تتلازم أو تتصاحب معه كلمة «فوات» وكلمة «الأوان» (قبل فوات الأوان) فإن هذا التركيب اللغوي ككل يتحول إلى تعبيرٍ اصطلاحي ليعطي معنى «العجلة» أو «الإسراع» في اتخاذ التدابير اللازمة. وغيرها الكثير من الأقوال المأثورة، والأمثال (Eddakrouri, 2016, p. 121).

٣. المفردات والنحو lexis and grammar:

يُنظر كثيرًا إلى اللغة على أنها تكاملٌ بين المفردات والنحو. أي أنه ليس هناك فرقٌ بين الكلمات المعجمية lexical words (أو ما يُطلق عليه أحيانًا كلمات المحتوى content words)، وبين الكلمات النحوية grammatical words (أو ما يُسمى أحيانًا الكلمات الوظيفية functional words أو الكلمات المفرغة empty words (Sinclair, 1991). غير إن الطرح الأكثر تداولًا في ضوء المدونات اللغوية هو أن هناك فارقًا كبيرًا بين المفردات والنحو. إذ إن الكلمات النحوية تكون عادةً أكثر ترددًا منها في الكلمات المعجمية في معظم النصوص. بالإضافة إلى ذلك فإن الكلمات النحوية ليس لها معنى دونما أن ترتبط بكلماتٍ معجميةٍ تكسبها المعنى ذاته. الأمر الذي يمكن التحقق منه باستخدام المدونات اللغوية.

وفضلًا عن ذلك فإنه يمكن استثمار المدونات اللغوية في هذا السياق في التحقق من ظاهرة التلازم اللغوي[1] collocation (كلمة محتوى + كلمة محتوى)، وظاهرة التلازم

١– يُسميها البعض أيضًا: تلازم المصطلحات، أو التلازم المعجمي، أو المصاحبة اللغوية، أو التلازم اللفظي. وهي كلماتٌ تتصاحب أو تتلازم أو تَقترن دائمًا بكلماتٍ أخرى، سواءٌ لسببٍ، أو دون سببٍ ظاهرٍ أو منطقيٍ.

المعجمي النحوي colligation (كلمة محتوى + كلمة وظيفية) من خلال استرجاع كافة سياقات الكلمات المعنيَّة (Hunston, 2002, pp. 149-51).

٤. التنوع اللغوي language variation:

يمكن الاستفادة من المدونات اللغوية في دراسة تنوع اللغة المستخدمة في مختلف المواقف، أو الفترات الزمنية، أو الأغراض، أو الظروف، أو الجماعات البشرية، أو المناطق الجغرافية، أو النوع (ذكر/ أنثى)، أو غيرها. ومن ثم استخدام هذه العوامل كونها متغيراتٍ تجريبيةً يمكن من خلالها التحقق من أكثر الكلمات ترددًا، و/ أو معاني الكلمات واستخداماتها words meanings and uses في كل متغيرٍ. ومن الممكن أن يَستتبع ذلك تحليلاتٍ لغويةٍ أخرى؛ كالتحقق من نوع وخصائص الجمل المستخدمة في كل متغيرٍ. وبالتالي يمكن الخروج بنتائج تعكس سمات التنوع في اللغة المستخدمة.
يمكن أيضًا الاعتماد على المدونات اللغوية في دراسة السجل اللغوي register أو النوع الأدبي genre؛ أي التحقق من لغة الخطاب discourse المستخدم في سياقاتٍ معينةٍ. كما هو الحال عند المقارنة بين السجل اللغوي المستخدم في نصوص الصحف، أو نصوص الدراسات الأكاديمية، أو أجزاء من النصوص العلمية، كالمقدمة، أو المستخلص، أو الخاتمة، أو غيرها.

هذا فيما يخص الإفادة من المدونات اللغوية من جانب الباحثين. أما فيما يخص ذلك من جانب الطلاب، فإنه يمكنهم الاعتماد على المدونات اللغوية في الإمداد بالمعلومات والكشف عن الحقائق بأنفسهم كونها نوعًا من التعلم الذاتي. مثل إمكانية اكتشافهم بأنفسهم لمعاني المفردات أو حروف الجر المستخدمة مع كلماتٍ معينةٍ، أو غيرها، من خلال السياق. وفي هذا الصدد يمكن أيضًا للمعلم أن يقدم السياقات المختلفة لكلمةٍ بعينها في شكل جملٍ، ويترك طلابه يكتشفون المعنى بأنفسهم. أو يطلب المعلم من طلابه أن يذهبوا إلى مدونةٍ لغويةٍ معينةٍ كي يبحثوا بأنفسهم ويتحققوا من المعنى.

وفضلًا عن ذلك فإنه يمكن استثمار المدونات اللغوية من جانب المعلمين في بناء وتصميم الدروس. إذ يمكن لهؤلاء المعلمين الحصول على سياقاتٍ جاهزةٍ لتقديمها كونها أمثلةً لتعزيز هدفٍ معينٍ، أو تضمينها في التدريبات والأسئلة اللغوية المختلفة (Hunston, 2002, pp. 137-97).

هذا، ويمكن الاطلاع على الدراسات التي وظَّفَت المدونات اللغوية بالفعل للإفادة منها في النحو والدلالة، وتدريس وتعلم اللغات، وذلك بالرجوع إلى الفصل الخامس من هذا الكتاب.

الإفادة من المدونات اللغوية في علم اللغة الاجتماعي Sociolinguistics

تُمثَّل المدونات اللغوية أحيانًا وفقًا للسن، أو للنوع، أو للمستوى التعليمي، أو للمستوى الاجتماعي، أو للمستوى الاقتصادي، أو للمكان الجغرافي... أو غيرها من العوامل والمتغيرات الاجتماعية التي يتم بها تمثيل representing النصوص في المدونات اللغوية. وكل هذه المتغيرات يمكن أن تقوم عليها دراساتٌ تقييميةٌ وبحوثٌ مقارِنةٌ من أجل التحقق من مدى تأثير هذه المتغيرات على المحتوى بوجهٍ عامٍ (O'Keeffe; McCarthy & Carter, 2007, p. 20).

فعلى سبيل المثال، قد يتم الاعتماد على مدونةٍ لغويةٍ لمقالات أو أخبار مجموعةٍ من الصحف العربية المنشورة ببلدانٍ عربيةٍ مختلفةٍ. ومن ثم يمكن التحقق من التنوع اللغوي في لغة الإعلام لهذه البلدان في ضوء المطابقة النحوية (Parkinson, 2008)، أو الإملاء (Zawaydeh & Saadi, 2006) واختلافاتها المتأثرة بالعوامل الجغرافية.

إذ قد ترتبط هذه الظواهر اللغوية بالأماكن والأقاليم الجغرافية التي يمكن للباحثين أن يُرْجِعوها إلى الآثار التي أحدثها الاستعمار في الوطن العربي. حيث تأثرت كل دولةٍ عربيةٍ بالدول الأخرى التي احتلتها، كإنجلترا، وفرنسا، وإيطاليا. الأمر الذي كان له أبلغ الأثر في إقحام الكثير من الكلمات المقترَضة (الكلمات الدخيلة loan words) من اللغات الأجنبية إلى اللغة العربية كي تجد لها الأرض الخصبة في لغة الإعلام (Abdelali, 2004).

هذا، ويمكن الاطلاع على الدراسات التي استثمرت بالفعل المدونات اللغوية للإفادة منها في علم اللغة الاجتماعي، وذلك بالرجوع إلى الفصل الخامس من هذا الكتاب.

الإفادة من المدونات اللغوية في صناعة المعاجم

أحدثت المدونات اللغوية ثورةً في عالم صناعة المعاجم حتى أضحت أداةً رئيسةً لصانعيها، لاسيما الضخمة منها. وربما تعود هذه الطريقة، التي تَعتمد فيها مداخلُ المعاجم على الاستخدام الفعلي للكلمات بصرف النظر عن الاعتماد على الحدس، إلى

القرن الثامن عشر الميلادي؛ ذلك حينما قام صامويل جونسون Samuel Johnson بإنشاء أول مدونةٍ لغويةٍ يدويةٍ لنصوصٍ ما بين عامي ١٥٦٠ و ١٦٦٠ م بغرض إعداد معجمٍ شاملٍ للغة الإنجليزية.

حيث تسهم المدونات اللغوية في رصد ومتابعة الكلمات الجديدة التي تدخل اللغة، وتحدد وقت دخولها، وتتحقق من الكلمات الموجودة بالفعل التي اكتسبت معنىً جديدًا. ويوجد الآن الكثير من معاجم اللغة الإنجليزية التي تحتوي على تواريخ مفصلةٍ لكل كلمةٍ، وأصلها اللغوي، ومتى تم نحتها أو استخدامها لأول مرة.

ويستطيع صانعو المعاجم بمساعدة المدونات اللغوية البحث في ملايين الكلمات والجمل، وسياقاتها المختلفة، واسترجاع كافة الأمثلة التي وردت بها من أجل تحري استخداماتها، وتحري المصطلحات والكلمات التي ترد قبلها و/ أو بعدها، ومن ثم يمكن تحديد التلازم اللغوي collocation، والتعبيرات الاصطلاحية idiomatic expressions بسهولةٍ. الأمر الذي يُسهم في تيسير سُبل الإفادة منها في الأغراض التعليمية، فضلًا عن أغراض استرجاع المعلومات (Dash, 2015).

ويلخص هانكس (Hanks, 2000) فوائد استثمار المدونات اللغوية في صناعة المعاجم من خلال تسليطه الضوء على ثلاثة اعتباراتٍ رئيسةٍ، هي:

١. لا يمكن الاعتماد فقط على تعريفات المعاجم للمعنى الحرفي للمفردات، وتكرار تردد الكلمات، والبديهة؛ بل ينبغي الركون باديء ذي بدءٍ على ما تقدمه المدونات اللغوية من بياناتٍ وملاحظاتٍ.

٢. تكشف البياناتُ التي تقدمها المدونات اللغوية الاختلاف في معاني الكلمات باختلاف تراكيبها النحوية، ومواقعها في السياق.

٣. توضِح غالبًا البياناتُ التي تقدمها المدونات اللغوية أن المعاني المجازية للكلمات تأتي أولًا، أو أنها تعكس المعنى الدارج للاستخدام الواقعي لها.

ومن أمثلة المعاجم التي لجأت إلى استخدام المدونات اللغوية كونها أداةً رئيسةً في صناعتها، معجم ديكشناري دوت كم Dictionary.com للغة الإنجليزية (2015)، ومعجم روبرت LE ROPERT للغة الفرنسية (2015)، ومعجم كولينز (Collins 2015) متعدد اللغات، ومعجم ماكميلان Macmillan للغة الإنجليزية (2015).

إضافةً إلى العديد من دور النشر التي أضحت تستخدم المدونات اللغوية في إنشاء معاجمها، مثل دار نشر جامعة أوكسفورد Oxford University Press، ودار نشر جامعة كمبريدج Campridge University Press. وكل هذه المعاجم تعتمد على مدونة سكيتش إنجين Sketch Engine (2015) في إنجاز مشروعاتها.

ومن الناحية البحثية، فإنه بمقدور مثل هذه المعاجم أن تجعل الباحثين أكثر ثقةً وتأكدًا من أن نتائج دراساتهم تعكس المعاني والاستخدامات الفعلية للكلمات إلى حدٍ كبيرٍ. إذ يقوم صانعو المعاجم بإعداد إحصاءاتٍ بتكرار تردد الكلمات، ومن ثم يمكن التمييز بدقةٍ بين ما هو مألوفٌ وما هو غريبٌ، وما هو محتملٌ وما هو ممكنٌ من الكلمات.

ويمكن توضيح المبادىء العامة التي تحكم صناعة المعاجم القائمة على المدونات اللغوية على النحو التالي:

١. لا تُفَضِّل المدونات اللغوية اللغةَ رفيعةَ المستوى:

تهدف المعاجم الإرشادية prescriptive lexicons عادةً إلى الحفاظ على نقاء اللغة، ولذلك فإنها تَستخدم مدونات لغوية لمؤلفي الوثائق ذوي المستوى الرفيع. إلا أن المعاجم الوصفية descriptive lexicons القائمة على المدونات اللغوية، وهي الأكثر شيوعًا، فينبغي أن تعكس الصورة الواقعية الصحيحة للغة التي تُستخدم بواسطة معظم متحدثيها.

٢. لابد أن تكون المدونات اللغوية كبيرةً ومتنوعةً:

ينبغي أن تغطي المدونات اللغوية المستخدَمة في المعاجم نطاقًا واسعًا من النصوص. ولا يوجد حدٌ معينٌ لحجم هذه المدونات اللغوية؛ فقد يتراوح حجمها بين أقل من مليون هيكل كلمةٍ إلى أكثر من مليار هيكل كلمةٍ. وقد أوضح قانون زبف Zipf's Law أن عددًا قليلًا من الكلمات يتكرر بشكلٍ كبيرٍ، بينما يندر تكرار الكثير من الكلمات؛ وهذا يعني أن اللغة تتكون من عددٍ قليلٍ من الكلمات الأكثر شيوعًا، وعددٍ كبيرٍ من الكلمات نادرة الاستخدام. وهذا يعني أيضًا أنه إذا ما أُريد التحقق من الكلمات نادرة الاستخدام أو قليلة التكرار، فإنه ينبغي أن يكون حجم المدونة اللغوية كبيرًا قدر الإمكان.

٣. المدونات اللغوية إما أن تكون متزامنةً أو تكون متعاقبةً:
ينبغي البت في تجميع النصوص من مراحل زمنيةٍ أو تاريخيةٍ مختلفةٍ (diachronic) قبل الشروع في إنشاء المدونة اللغوية، كما هو الحال في مدونة هيلسينكي للنصوص الإنجليزية Helsinki Corpus of English Text (2011). وقد يتم تجميع النصوص من فترةٍ زمنيةٍ معاصرةٍ (synchronic)، كما هو الحال في المدونة اللغوية الدولية للغة الإنجليزية The International Corpus. of English (ICE) (2015). ومما هو مؤكدٌ أن المعاجم التاريخية تتطلب مدوناتٍ لغويةً متعاقبةً، بينما تتطلب المعاجمُ التعليمية مدوناتٍ لغويةً متزامنةً توضح كيفية الاستخدام الراهن للغة.

٤. يُفضل أن تكون المدونات اللغوية متوازنةً:
لا يمكن للمدونات اللغوية أن تتبع الأساليب المنهجية في تجميع العينات العشوائية؛ لأن موضوع العينة هنا هو النصوص التي هي بالأساس عبارةٌ عن وحداتٍ لغويةٍ ديناميكيةٍ تنمو وتتطور. فديناميكية اللغة تَحُول دون الفهم الكامل لطبيعة وحدود اللغة. ولذا فإن ما تطمح إليه معالجة النصوص هو إنشاء مدوناتٍ لغويةٍ متوازنةٍ من خلال الاعتماد على أسلوب العينة الطبقية. حيث يتم تصنيف النصوص في فئاتٍ تبعًا لأنواعها وموضوعاتها، ومن ثم يتم الحصول على عينةٍ عشوائيةٍ من هذه الفئات. ولإضفاء أكبر قدرٍ من التوازن على النصوص المُجَمَّعَة، فإنه لا يُكتفى بوضع أنواع النصوص في الحسبان فحسب، بل أيضًا نسب أو عزو هذه النصوص بعضها إلى بعض (Attia; Tounsi & Genabith, 2010).

٥. ينبغي أن تكون المدونات اللغوية موضوعيةً:
هذا يعني أنه يجب أن تكون النصوص التي يتم جمعها غير متحيزةٍ لفئةٍ أو نوعٍ معينٍ من النصوص على حساب غيره؛ حتى لا تَفتقد النتائج المسترجعة إلى الدقة والموثوقية. فمثلاً إذا كان الهدف من المدونة اللغوية هو تحليل مقالات الأخبار أو الأعمال الشعرية، فإنه لا يمكن بحالٍ أن يتم تعميم النتائج على كافة أنواع النصوص الأخرى (Attia; Tounsi & Genabith, 2010).

هذا، ويمكن الاطلاع على الدراسات التي استثمرت بالفعل المدونات اللغوية للإفادة منها في صناعة المعاجم، وذلك بالرجوع إلى الفصل الخامس من هذا الكتاب.

الإفادة من المدونات اللغوية في الترجمة

للمدونات اللغوية تطبيقاتٌ عدةٌ في مجال الترجمة؛ فالبعض يستخدم المدونات اللغوية التي تحوي الترجمات ويقارنها بالمدونات اللغوية التي تحتوي على النصوص الأصلية؛ بهدف التحقق من الخصائص المميزة للنصوص المترجمة (Baker, 1995, pp. 223-25). بينما يستخدمها البعض الآخر عاملًا مساعدًا في عملية الترجمة ذاتها، سواءٌ أكانت يدويةً أم آليةً. والمدونات اللغوية في هذا السياق تنقسم إلى ثلاثة أنواعٍ:

١. المدونات اللغوية أحادية اللغة monolingual:

تتكون من لغةٍ واحدةٍ، سواءٌ أكانت لغة النصوص الأصلية، أو لغة النصوص المترجمة.

٢. المدونات اللغوية المقارِنة comparable:

تحوي نصوصًا من لغتين (أو أكثر) مختلفتين (كالعربية والإنجليزية على سبيل المثال)، أو من نوعي لغة two varieties (أو أكثر) مختلفين (كالفصحى والعامية). بحيث يحتوي كلا منها على القدر نفسه من النصوص، بما يسمح بالمقارنة الموضوعية والعادلة بينهما.

ويمكن أن تُستخدم المدونات اللغوية المقارِنة لنوعي لغةٍ مختلفين من أجل المقارنَة بينهما. بينما تُستخدم المدونات اللغوية المقارِنة للغتين مختلفتين لأغراض الترجمة وتدريس اللغات؛ بحيث يمكن التحقق من أوجه التشابه ونقاط الاختلاف بين هذه اللغات.

وتُعد اللغة العربية من اللغات الفقيرة في هذا الصدد. ومن أبرز الأمثلة على هذا النوع المدونة اللغوية الدولية للغة الإنجليزية The International Corpus of English (ICE) (2015) التي تضم أنواعًا مختلفةً من الإنجليزية.

٣. المدونات اللغوية المتوازية parallel:

تتكون من نصوصٍ أصليةٍ في إحدى اللغات جنبًا إلى جنبٍ مع نظيرتها المترجَمة في لغةٍ أخرى واحدةٍ أو أكثر. وهي إما أن تكون أحادية الاتجاه unidirectional من اللغة أ إلى اللغة ب، أو ثنائية الاتجاه bidirectional من اللغة أ إلى اللغة ب، ومن اللغة ب إلى اللغة أ (Aston, 1999).

ويمكن استثمار هذا النوع من المدونات اللغوية في إرشاد المترجمين والطلاب إلى الكلمات والتعبيرات اللغوية المقابِلة بين اللغات، ومن ثم إمكانية المقارنة بينها (Hunston, 2002, pp. 14-15).

مثالٌ على ذلك القاموسُ متعدد اللغات على الإنترنت المعروف باسم: Glosbe الذي يترجم من وإلى الكثير من اللغات الحية، ومنها اللغة العربية.

هذا، ويمكن الاطلاع على الدراسات التي استثمرت بالفعل المدونات اللغوية للإفادة منها في الترجمة، وذلك بالرجوع إلى الفصل الخامس من هذا الكتاب.

الإفادة من المدونات اللغوية في دراسة التوجهات الفكرية (الأيديولوجيا)

هناك اهتمامٌ متزايدٌ بدراسة العلاقة بين اللغة والأيديولوجيا، ولاسيما فيما يتعلق بدور اللغة في تشكيل الافتراضات حول العالم المحيط، ونقل الإرهاصات عما يمكن أن تكون عليه الشؤون الدنيوية. إذ يُنظر إلى اللغة في هذا السياق على أنها نسيجٌ متغلغلٌ في العالم الاجتماعي يُسهم في تخليد هذه الافتراضات والإرهاصات والقيم حول هذا العالم الاجتماعي (Fowler, 1987).

ويمكن هنا استخدام المدونات اللغوية في ضوء ثلاثة محاور رئيسةٍ:

١. دراسة النصوص من خلال التحقق من سياقات الظروف الاجتماعية التي نشأت فيها.

٢. الكشف عن الأيديولوجيات الضمنية وراء الافتراضات المعلَنة.

٣. التأكد من المغزى العام للأشياء التي يتم تمثيلها بمدلولاتٍ مختلفةٍ.

وتتضح استخدامات المدونات اللغوية جليًا في المحورين الثاني والثالث. حيث يمكن الإفادة من المدونات اللغوية في هذا الإطار في التحقق من التركيبة المعجمية للمفردات مع بعضها البعض داخل النصوص كما وردت في اللغة الطبيعية؛ بهدف الكشف عن الأيديولوجيات أو المغزى من الرسائل الضمنية أو الغريبة المنقولة بعباراتٍ صريحةٍ.

غير إن هناك بعض الشكوك وراء مدى إفادة المدونات اللغوية فيما يخص المحور الأول. إذ يرى البعض أنه بمجرد معالجة نصوص المدونات اللغوية فإنه بذلك تكون قد خرجت من لغتها أو سياقاتها الطبيعية، ومن ثم لا يمكن الحكم على العوامل الاجتماعية المحيطة أو المؤثرة فيها.

وعلى أية حالٍ فإن المدونات اللغوية تسهم بشكلٍ كبيرٍ في التحقق من الخلفية الأيديولوجية للمؤلفين من خلال تحليل الخطاب discourse analysis. فعلى سبيل المثال، يمكن الاستعانة بالمدونات اللغوية في تعيين الكلمات المفتاحية لأي مؤلفٍ يكتب

في الأمور السياسية أو الاجتماعية أو الثقافية... إلخ. حيث يتم فحص هذه الكلمات التي يستعين بها باستمرارٍ في كتاباته، وهي في الوقت نفسه تُوظَف من أجل خدمة توجهاته الفكرية، على حساب كلماتٍ مفتاحيةٍ أخرى قد لا يؤمن بها هذا المؤلف (Hunston, 2002; Tuebert, 2000).

الإفادة من المدونات اللغوية في علم المعلومات

لطالما استَخدَم علمُ المعلومات مصطلحَ «مجموعة اختبار» test collection للدلالة على قياس كفاءة وصلاحية الأدوات، ودقة وفعالية الطرق والنظم المستخدمة في تجارب استرجاع المعلومات. وكانت البداية مع تجارب كرانفيلد Cranfield experiments في منتصف خمسينيات القرن الماضي. حيث استُخدِم هذا المصطلح حينها ليشير في حقيقته إلى:

١. مجموعةٍ من الوثائق التي تشتمل على العناوين، أو المستخلصات، أو النصوص الكاملة، أو كل هذه الأجزاء معًا.
٢. مجموعةٍ من الاستفسارات أو الأسئلة.
٣. مجموعةٍ من أحكام الصلاحية الصادرة بحق هذه الوثائق (MacMullen, 2003).

ومن بين أبرز مجموعات الوثائق التي تم استخدامها في تجارب استرجاع المعلومات «مجموعات كرانفيلد» (Salton, 1971) ومجموعة اختبار وكالة أنباء رويترز Reuters (Sanderson, 1994)، ومجموعات تريك TREC collections (TREC) Text REtrieval Conference, National Institute of Standards and Technology (NIST), 2015.

إلا أن هذه التجارب واجهتها مشكلاتٌ تطبيقيةٌ تمثلت في أن الحكم على صلاحية الوثائق لا يعني بالضرورة أنها تلبي احتياجات الباحثين الواقعية كليةً. إذ إن هناك جانبًا غايةً في الأهمية أغفلته هذه التجارب، وهو إلى أي مدى تمثل هذه الوثائق الواقع الفعلي real-life للنصوص كما يستخدمها المستفيدون. ويُضاف إلى ذلك أن حجم هذه الوثائق لم يكن من الشمولية والتغطية التي يمكن معهما القول بأنه يمثل الواقع الذي استخدمت فيه. وبالتالي ظل التصديق بصلاحية هذه النظم في التطبيق منقوصًا (Ledwith, 1992, p. 452).

ولذا أصبحت المدونات اللغوية بمفهومها الشمولي الحديث أداةً لا غنىً عنها في اختبار وتقييم نظم استرجاع المعلومات، على النحو الذي يعكس الاستخدام الحقيقي لمدخلات هذه النظم. إضافةً إلى استخدامها في الكشف عن المعرفة في قواعد البيانات Knowledge Discovery in Databases (KDD)، والتنقيب عن البيانات data mining، واقتباس/استخراج المعلومات information extraction، والتكشيف الدلالي الكامن latent semantic indexing .

كما يمكن أن تُستخدم المدونات اللغوية كونها أداةً للتحليل الكمي والنوعي في علم المعلومات. فعلى سبيل المثال، يمكن تجميع نصوص مجموعةٍ من المقالات المتخصصة في علم المعلومات، ويتم التحقق من عدد مرات تكرار كلمةٍ مثل «معلومات» أو «مكتبات». حينها نحصل على نتائج كمية تُمكِّن من معرفة أي الكلمات أكثر تكرارًا، وأيها أقل. ومن ثم نستطيع أن نَخلُص إلى أن إحدى الكلمات أكثر تداولًا، وبالتالي فهي معتمَدةٌ أكثر من غيرها. وفي الوقت نفسه نستطيع أن نتحقق من الشكل الذي وردت به إحدى الكلمات أو العبارات أو الجمل. أي أن المدونات اللغوية لا تقف عند حدود التحليل الكمي فحسب، بل إنها تقوم أيضًا بتحليلٍ نوعيٍ يُظِهر البيئة اللغوية الداخلية وكذلك البيئة المعرفية المحيطة بالمادة المسترجَعة.

وجديرٌ بالذكر أنه لا توجد مدونةٌ لغويةٌ تَصْلَح للإفادة منها في كافة الأغراض. فقد يستخدمها البعض لأغراضٍ بحثيةٍ في دراسة المفردات والمصطلحات وأشكالها. وقد يستخدمها البعض الآخر في المقارنة بين المفردات والمصطلحات في اللغات واللهجات المختلفة. مثل المقارنة بين مصطلح «metadata» في الإنتاج الفكري الأجنبي، ومقابلاته العربية المختلفة التي استخدمها الباحثون. وفي هذه الحالة لابد أن تُمثِل المدونةُ اللغويةُ الإنتاجَ الفكري تمثيلاً دقيقًا. وقد تفي بعض المعاجم أو نظم استرجاع المعلومات بقدرٍ من هذه الأغراض، إلا أن البحث في المدونات اللغوية يوفر أمثلةً كثيرةً ومتنوعةً لكلمات البحث في سياقاتها الطبيعية (MacMullen, 2003).

وبوجهٍ عام، فإنه يمكن الإفادة من المدونات اللغوية في علم المعلومات في ضوء:

١. التحقق من موقع أي عنصرٍ من عناصر النص (اسم، أو فعل، أو حرف، أو جملة، أو عبارة، أو مقطع، أو فقرة... إلخ) حتى وإن كان هذا العنصر كلمةً واحدةً فقط، أو ماثلًا في ذاكرة المستفيد جزئيًا، فإنه بالإمكان العثور عليه وتسجيل مصدر الاقتباس كاملًا.

٢. مقارنة وتحليل الكلمات، وخاصةً المشتركات اللفظية، والمترادفات، والأشكال المختلفة للمصطلحات. حيث يمكن حصر كل المعاني المختلفة لكلمةٍ بعينها من خلال عرض كافة سياقاتها، مثل كلمة «قلب» التي من معانيها: جوهر، وقلب يضخ الدماء، ووسط أو منتصف. أما إذا ما تم حصر المترادفات المختلفة لكلمةٍ معينةٍ، فإنه يمكن التحقق من معدل شيوع كلٍ منها، مثل كلمات: ميتاداتا، وميتاديتا، وما وراء البيانات، والبيانات الفوقية، وبيانات عن البيانات... أو مصادر المعلومات، وأوعية المعلومات، ومواد المعلومات... أو تكنولوجيا المعلومات، وتقنيات المعلومات... إلخ.

وكما نلاحظ أن أول كلمتين تمثلان شكلين مختلفين لمصطلحٍ واحدٍ. وبالتالي يمكن إرشاد أو إحالة الباحثين إلى أي المصطلحات يمكن استخدامها، وأيٍ منها يمكن إغفاله.

٣. التأكد من مظاهر البحث والتطور في أي مجالٍ من مجالات المعرفة، ومجالات الاهتمام الحديثة في أحد التخصصات، سواءٌ بوجهٍ عامٍ، أو لمؤلفٍ بعينه عبر مراحل زمنيةٍ مختلفةٍ.

٤. فحص الاختلافات والتشابهات في استخدام الكلمة بين المؤلفين المختلفين، وأيضًا في الكتابات المبكرة والأخيرة لمؤلفٍ بعينه. الأمر الذي من شأنه أن يُسهم في التحقق من التأليف وصناعة الكتابة والانتحال (عبد الهادي، ١٩٨٢، ص ٢٧، ص ص ٧١-٧٢).

٥. مقارنة واستعمال الكلمات في دراسات الأسلوبية stylistics؛ إذ تتيح المدونات اللغوية التحقق من أوجه التشابه ومظاهر الاختلاف في استعمال الكلمات من جانب المؤلفين. ويمكن لمثل هذا التحليل أن يساعد في التأكد من مسؤولية تأليف الأعمال مجهولة المؤلف (قاسم، ٢٠٠٠، ص ص ١٩-٢٣).

هذا، ويمكن الاطلاع على الدراسات التي استثمرت بالفعل المدونات اللغوية للإفادة منها في استرجاع المعلومات، وذلك بالرجوع إلى الفصل الخامس من هذا الكتاب. كذلك يمكن مطالعة الدراسات التي وظفت المدونات اللغوية للإفادة منها في المكتبات وعلم المعلومات بالرجوع إلى الفصل نفسه.

الإفادة من المدونات اللغوية في صناعة المكانز

هناك شقان رئيسان لدراسات المصطلح القائمة على المدونات اللغوية، هما: رصد المصطلحات؛ أي اكتشاف المصطلحات المرشحة، والإنشاء الآلي للمكانز؛ أي إضافة

علاقاتٍ دلاليةٍ لبنك المصطلحات الذي يتم تجميعه (Morin & Jacquemin, 1999, pp. 389-93).

ولطالما اعتمد الباحثون على الطرق اليدوية في إنشاء المكانز، غير إن هذه الطرق تكتنفها مشكلتان رئيستان، هما: حجم الجهود اليدوية الكبيرة المبذولة في تحديد المصطلحات التي يتم إدراجها وتنظيمها داخل المكنز، ومدى ملاءمة أو تغطية المكنز اليدوي هذا لمجموعات الوثائق. علاوةً على ظهور هاتين المشكلتين مرةً أخرى إذا ما أُريد تحديث المكنز. ولذا عكف الباحثون على إيجاد البدائل التي من شأنها تسريع وتيرة إنشاء وتحديث المكانز باللجوء إلى برمجيات معالجِة لمحتوى المدونات اللغوية (Grefenstette, 1993).

وربما تعود فكرة استخدام الحاسب الآلي في إنشاء المكانز إلى سبعينيات القرن الماضي حينما حاول سالتون (Salton, 1971) وجونز (Jones, 1971) استخدام مصفوفاتٍ رياضيةٍ توضح العلاقة بين المصطلحات من خلال مقاييس إحصائيةٍ. إلا أن هذه الطريقة تكتنفها بعض المشكلات أيضًا؛ وهي أن الكثير من المصطلحات غير المرتبطة ببعضها البعض تتصاحب معًا كونها عامةً أو كثيرةً في تكرار ترددها، كما أنه من الصعوبة بمكانٍ استخدام المترادفات معًا، وأن اعتماد المصطلحات المكونة من كلمةٍ واحدةٍ يأتي على حساب المصطلحات المكونة من أكثر من كلمةٍ، إضافةً إلى أن التجمعات العنقودية clusters للمصطلحات المترابطة ترد عادةً دون إظهار العلاقة بين وحداتها (Jing & Croft 1994, pp. 146-55).

إلا أنه في أواخر تسعينيات القرن المنصرم أتاحت بعض برمجيات معالجة المدونات اللغوية طُرقًا للتحليل القواعدي للنصوص لتوليد أو إنشاء المكانز بفرضية أن المصطلحات المتشابهة أو المترادفة تظهر عادةً في علاقاتٍ نَظْميةٍ واحدةٍ، حيث يتم تجميعها معًا وفقًا للسياق النحوي أو الآجرومي الذي وردت به (Grefenstette, 1994, pp. 18-23).

وقد حدد البعض خطوات إنشاء المكانز على النحو التالي:

١. إعداد مدونةٍ لغويةٍ.

٢. تحديد سياقات كل كلمةٍ.

٣. تعيين الكلمات التي تشترك في السياقات نفسها.

وبوجهٍ عامٍ، فإن دقة المكانز تزداد بازدياد حجم المدونات اللغوية القائمة عليها؛ فكلما ازداد حجم المدونة اللغوية، ازدادت القدرة على تعيين الكلمات المتشابهة وتمييزها عن الكلمات الأخرى المشوِشة (Rychly & Kilgarriff, 2007, pp. 41-44).

ومن أمثلة المكانز التي تعتمد على المدونات اللغوية كونها أداةً رئيسةً في صناعتها، مكنز Thesaurus.com (2015)، ومكنز كولنز Collins للغة الإنجليزية (2015)، اللذان يستخدمان مدونة سكيتش إنجين Sketch Engine (2015) في إثرائهما.

الإفادة من المدونات اللغوية في المعلوماتية الحيوية Bioinformatics

أصبحت المدونات اللغوية أداةً رئيسةً في المعلوماتية الحيوية وعلم الأحياء بوجهٍ عام. حيث يتم الاعتماد عليها في المقارنة بين الجينومات والمنتجات البروتينية للجينات؛ من خلال التحقق من تسلسل الـ دي إن إيه DNA، والـ آر إن إيه RNA، والبروتينات ومضاهاتها أو مقارنتها في الكائنات الحية. وبالتالي يمكن تحديد درجات التشابه، والاختلاف، والوقت المستغرق عبر مراحل التطور المختلفة (Tanabe & et al. 1999). ومن أمثلة المدونات اللغوية المستخدَمة في توقع التركيبات البروتينية مشروعُ تقويم تحشية الجينوم Genome Annotation aSsessment Project (GASP) (2014)، والمدونة اللغوية ذات التحشية متعددة الأوجه للعلاقات بين الجينات ومرض السرطان Corpus with Multi-faceted Annotations of Gene-Cancer Relations (CoMAGC) (Lee & et al. 2013).

الإفادة من المدونات اللغوية في المعلوماتية الجنائية Forensic Informatics

يُقصد بالمعلوماتية الجنائية تطبيقات نظم المعلومات وعلم اللغة في مجال القانون، ولاسيما فيما يخص القضايا الجنائية والأحول المدنية على السواء. إذ تُستخدم المدونات اللغوية كونها أداةً رئيسةً في معالجة محاضر تحريات الشرطة، والتأكد من هُوِيات المتحدثين، ونصوص القوانين والمرافعات، ومدى قابلية النصوص التشريعية للقراءة والفهم، ومصداقية الوثائق التي تتضمن اعترافات ضحايا حالات الانتحار، وخطابات التهديد.

كما يمكن أن تفيد المدونات اللغوية في التحقق من الأصوات المسجلة، وما إذا كانت وثيقتان (أو أكثر، أو فقرات من الوثائق) منسوبتين لشخصٍ ما، أو لمجموعةٍ من الأشخاص. ويمكن أن تفيد أيضًا في الحكم على مدى مصداقية أقوال المتهمين. فما يتم

تجميعه من نصوصٍ في هذا الشأن يمكن أن يستخدم في بناء مدونةٍ لغويةٍ يتم معالجتها على يد المتخصصين. الأمر الذي من شأنه أن يسهم في تقديم الأدلة، ومن ثم إصدار الأحكام اعتمادًا على البراهين والإثباتات.

وعلى النحو نفسه يمكن أن تُستخدم المدونات اللغوية في التحقق من سرقة وانتحال المعلومات العلمية plagiarism. حيث تُستخدم برمجياتٌ مخصصةٌ لهذا الغرض في تحديد النصوص المنسوخة؛ من خلال مقارنة النصوص المشكوك فيها suspected texts بتلك النصوص الضابطة control texts. ومع وضع حدٍ أدنىً لعدد المفردات المشتركة بينهما، والذي يتم معه القبول بسلامة النصوص المنسوخة، فإنه يمكن الحكم على السرقة العلمية من عدمها. كما يمكن للمقارنة أن تتم في ضوء تحديد المفردات التي تكررت مرةً واحدةً فقط (يطلق عليها بالإنجليزية hapaxes) أو تلك المفردات الفريدة unique التي تتفرد بها مجموعة النصوص المشكوك فيها مقابل النصوص الضابطة، أو العكس. ومن الممكن أن تتم هذه المقارنة على مستوى العبارات أيضًا.

ومن أمثلة المدونات اللغوية المستخدَمة في المعلوماتية الجنائية مستودع المعلوماتية الجنائية للقياسات الحيوية The Forensics Informatics Biometrics Repository FIB-R) (2015) (Hunston, 2002, pp. 130-31; Suchomel & Brandejs, 2014; .(Garfinkel, 2009

مراحل إنشاء المدونات اللغوية

تمر أي مدونةٍ لغويةٍ بمجموعةٍ من مراحل الإنشاء التي يمكن توضيحها على النحو التالي:

تحديد الهدف والأساس المنطقي rationale

يُشترط في المدونة اللغوية أن تكون ممثِلةً لمجتمع أو شيءٍ ما؛ كأن تكون ممثِلةً لأحد فروع المعرفة، أو أحد المؤلفين، أو أحد الأماكن... وغير ذلك من الأمور التي يجب مراعاتها عند تصميم، أو شراء، أو استخدام المدونات اللغوية. ومعيار التمثيل representativeness هنا يؤدي دورًا رئيسًا في تحديد الأساس المنطقي عند إنشاء وتصميم المدونات اللغوية.

تحديد أنواع النصوص المناسبة

يراعى هنا التحقق من اعتبارات جدوى المدونة اللغوية؛ أي ما الذي يمكن إتاحته؟ وما هو مناسبٌ أخلاقيًا واجتماعيًا؟ مع مراعاة أن يكون تمثيل البيانات دقيقًا ومحقِقًا للغرض من المدونة اللغوية. ويراعى أيضًا تحديد ما إذا كانت النصوص مكتوبةً، أم منطوقةً، أم مزيجًا بينهما بعد تحديد نسب كل نوعٍ وفق أسسٍ مسبقةٍ.

تحديد مصادر النصوص

قد يلجأ البعض إلى بناء مدونةٍ لغويةٍ من خلال مقالات الدوريات، أو نصوص الكتب، أو الرسائل الجامعية... إلخ. أو أن يتم تسجيل كافة التقارير الإذاعية الإخبارية لإحدى الصحف، أو تسجيل التقارير الرياضية، أو التقارير الاقتصادية، أو السياسية... إلخ.

الحصول على الموافقة القانونية

يتعين قبل أي إجراءٍ على النصوص أن تُراعى أولًا حقوق التأليف والملكية الفكرية. أما إذا ما أُريد بناء مدونةٍ لغويةٍ منطوقةٍ فإنه يُفضل الحصول على موافقةٍ كتابيةٍ من صاحب النص قبل الشروع في عملية التسجيل، بحيث تُوضَح في هذه الموافقة الطريقةُ التي سيتم بها تسجيل الحديث، والغرض من ذلك.

جمع النصوص

حينما تكون المدونة اللغوية ممثِلةً لنصوصٍ مكتوبةٍ، فإنه يتم عادةً إدخالها إلى الحاسب الآلي عن طريق:

١. الأسلوب اليدوي التقليدي باستخدام لوحة المفاتيح.

٢. المسح الضوئي scanning باستخدام قارئ المحارف البصرية optical character reader OCR الذي يقوم بمسح النصوص المطبوعة ثم تحويلها إلى نصوصٍ رقميةٍ.

إلا أن ثمة مشكلتين تواجها هذه الوسيلة، هما:

- اقتصار تعامل برامج قراءة المحارف البصرية على أنماطٍ أو أشكالٍ معينةٍ من الحروف.
- غموض النصوص المطبوعة؛ الأمر الذي يؤدي إلى حدوث الكثير من الأخطاء في عملية قراءة المحارف. إلا أنه يمكن التغلب على هذه المشكلة بمراجعة النصوص عقب قراءتها ضوئيًا.

٣. التحميل مباشرةً من الشبكة العنكبوتية. حيث يمكن اعتبار الشبكة العنكبوتية مدونةً لغويةً شاملةً؛ نظرًا لما تحويه من مليارات الكلمات والنصوص التي يمكن استخدامها في البحث والتحليل (Kilgarriff & Grefenstette, 2003).

أما إذا كانت المدونة اللغوية ممثِلةً لنصوصٍ منطوقةٍ، فإن ذلك يستنفد الكثير من الوقت؛ لما يتطلبه من تسجيل النصوص الشفوية ونسخها transcripting. ويتم نسخ النصوص بإحدى هذه الطرق:

١. النسخ اليدوي: فقد قدر البعض أن ساعةً من الكلام المُسجَل تتراوح بين ١٠٠٠٠ إلى ١٥٠٠٠ كلمةٍ تأخذ نحو يومين لنسخها، وهذا يتوقف بالطبع على نوع الحديث.

٢. النسخ الآلي: يمكن نسخ النصوص باستخدام برامج الإملاء الآلي (كلامٌ إلى نصوصٍ speech-to-text).

٣. النسخ الصوتي:

- باستخدام رموزٍ بديلةٍ للألفباء الصوتية phonetic alphabet وهذا مطلوبٌ في تخزين النصوص العامية، أو اللهجية، وكذلك لغة الطفل.
- أو بإضافة رموزٍ للدلالة على منعطفات الحديث من إيماءاتٍ، وتداخلاتٍ لفظيةٍ، وصفاتٍ أخرى غير لفظيةٍ، ووقفٍ، وترددٍ... وما شابه ذلك من ظواهر تتسم بها النصوص المنطوقة والحوارات. ويوضح ذلك الجدول رقم (١) (Garg; Marti-novski & Robinson, 2004).

الجدول رقم (١) الأكواد المستخدَمة في المدونات اللغوية المنطوقة ووظائفها

الكود	الوظيفة
<$1> <$2> <$3>	للإشارة إلى كل متحدثٍ وفقًا لترتيب سماع أصوات كل منهم.
+	للإشارة إلى وجود مقاطعةٍ أثناء الحديث؛ حيث توضع علامتان؛ إحداهما في بداية المقاطعة، والأخرى بعد انتهاء المقاطعة والاستطراد في الحديث.

الكود	الوظيفة
=	للإشارة إلى الكلمات المبتورة truncated. مثل: متطلبا=
<?>	للإشارة إلى الكلام غير المفهوم. مثل: هذا يعني أن <?>
<SE> صفير <E$\>	للإشارة إلى المعلومات اللغوية الإضافية. مثل: نباح كلبٍ، ضحكةٌ، تثاؤبٌ، صوت سيارةٍ مسرعةٍ... والتي تتحدد مدى أهميتها وفقًا للغرض من المدونة اللغوية.

التحقق من النصوص التي تم جمعها

قد تتداخل طرق جمع النصوص مع بعضها البعض في الكثير من الأحيان. وفي كل الحالات لابد من مضاهاة ومراجعة وتدقيق النصوص جيدًا؛ الأمر الذي يَزيد من الوقت والجهد المستنفدين في إنجاز عملية جمع النصوص، ولاسيما المنطوقة منها. ولذا يضطر عادةً القائمون على المدونات اللغوية المنطوقة إلى تقليل حجم العينة للحد الأدنى الذي يسمح معه بتمثيل المجتمع؛ توفيرًا للوقت والجهد. ولذا نجد أيضًا أنه مهما بلغ حجم المدونات اللغوية المنطوقة فإنه يظل عادةً أقل من نظيره المكتوب (O'Keeffe & Farr, 2003, pp. 390-95).

حفظ النصوص في ملفات

يتم حفظ النصوص عادةً، سواءٌ أكانت مكتوبةً أو منطوقةً، في الشكل البسيط plain format لأنه الشكل الأنسب والأكثر مرونةً في التعامل مع مختلف البرمجيات. كما أنه يتم عادةً ضغط ملفات النصوص بعد حفظها لتوفير المساحة التي تشغلها على القرص الصلب Hard Desk (Wynne, 1997).

تشفير النصوص

تواجه الكثير من لغات النصوص المكتوبة بعضَ المشكلات في تشفيرها، ولاسيما التمثيلات الخاصة special characters منها. ولذا طُرحت بعض الحلول، منها إيجاد

مواصفاتٍ معياريةٍ تحكم عملية تشفير التمثيلات في المدونات اللغوية، مثل مواصفة الآيزو ISO لتشفير التمثيلات، ونظام الشفرة الموحدة Unicode (Pinkas, 2014). الأمر الذي من شأنه أن يُيسر عمليات المعالجة التي ستتم على هذه النصوص لاحقًا.

هذا، ويمكن الاطلاع على الدراسات التي تناولت كيفية إنشاء وإتاحة المدونات اللغوية، ولاسيما العربية منها، بالرجوع إلى الفصل الخامس.

أنواع المدونات اللغوية

ثمة أنواعٌ مختلفةٌ للمدونات اللغوية؛ فهناك من يقسمها تبعًا لطريقة معالجة نصوصها إلى: مدونات لغويةٍ مَحْشوَّةٍ annotated أو مدونات لغويةٍ مُرَمَّزةٍ marked-up، ومدونات لغويةٍ خامٍ raw.

وهناك من يقسمها تبعًا للغات النصوص التي تحويها إلى مدونات لغويةٍ أحادية اللغة، ومدونات لغويةٍ مقارِنةٍ، ومدونات لغويةٍ متوازيةٍ.

ويقسمها البعض الآخر تبعًا للغرض من استخدامها إلى نوعين:

مدونات لغويةٌ اختباريةٌ Test Corpora

عبارةٌ عن رصيدٍ من النصوص الأصلية أو المُخَلَّقة invented التي تُستخدم في اختبار، أو تجريب، أو تقييم، أو تقييس الأداء. وتسمى أيضًا مجموعات الاسترجاع التجريبية، أو مجموعات الاختبار test suits/collection.

مدونات لغويةٌ بحثيةٌ Research Corpora

عبارةٌ عن رصيدٍ من النصوص الأصلية التي تُستخدم في إجراء تجارب بحثيةٍ من أجل تطوير المعرفة؛ حيث تُستخدم كونها قاعدةً للتحليل الفكري بصفتها مستودعًا للغة الطبيعية. وينقسم هذا النوع إلى أربعة أنواعٍ من المدونات اللغوية:

مدونات لغويةٌ عامةٌ General Corpora

تضم أكبر قدرٍ ممكنٍ من النصوص في لغةٍ ما، على اختلاف أنواعها types، سواءٌ أكانت نصوصًا مكتوبةً أو نصوصًا منطوقةً، أو كلتيهما. وتُستخدم المدونات اللغوية العامة غالبًا كونها مصادر مرجعيةً في إنشاء كتب قواعد اللغة، والقواميس، وغيرها؛ ولذا يُطلق عليها أحيانًا المدونات اللغوية المرجعية. مثل: المدونة اللغوية الوطنية

البريطانية British National Corpus (BNC) (2015) التي تشتمل على أكثر من ١٠٠ مليون هيكل كلمةٍ.

مدونات لغويةٌ متخصصةٌ

تحوي نصوصًا من نوعٍ معينٍ (مقالاتٌ أكاديميةٌ، أو محاضراتٌ، أو مقالاتٌ تحريريةٌ... إلخ). وقد يكونُ هذا النوع من النصوص محددًا بفترةٍ زمنيةٍ، أو بمستوىً اجتماعيٍ، أو بتخصصٍ موضوعيٍ. ويتسم هذا النوع بأنه أصغر حجمًا مقارنةً بالمدونات اللغوية العامة.

ويقوم عادةً الباحثون أنفسهم بإعداد هذا النوع من المدونات اللغوية من أجل التحقق من نصوصها. وهم في ذلك غير مقيدين بدرجةٍ محددةٍ من التوغل في التخصص. وما يقيدهم فقط هو نوع النصوص التي يقومون بتضمينها. كأن يقوم باحثٌ مثلًا بتجميع نصوص مقالات الدوريات العلمية في تخصصه، دون نصوص الرسائل العلمية، أو غيرها من أنواع مصادر المعلومات الأخرى.

ومن أقرب الأمثلة لهذا النوع المعجمُ المفهرس لألفاظ القرآن الكريم التابع لمجمع الملك فهد لطباعة المصحف الشريف.

مدونات لغويةٌ تاريخيةٌ Historic **أو تعاقبيةٌ** Diachronic

وهي تحوي نصوصًا تنتمي إلى فتراتٍ زمنيةٍ معينةٍ؛ بهدف التحقق من تطور الكلمات والمصطلحات عبر فتراتٍ زمنيةٍ محدَدةٍ. وهذا النوع من المدونات اللغوية يجب أن يُشكِل الأساس لأي عملٍ معجميٍ تاريخيٍ يعتمد على الاستقراء المنهجي. ومن أمثلتها مدونة هيلسينكي للغة الإنجليزية Helsinki Corpus of English Texts التي تضم نصوصًا إنجليزيةً قديمةً وحديثةً (٢٠١٤). والمدونة الممثِلة للسجلات اللغوية الإنجليزية التاريخية A Representative Corpus of Historical English Registers (ARCHER) التي تشتمل على نصوصٍ للإنجليزية البريطانية والأمريكية للفترة ما بين عامي ١٦٠٠ و ١٩٩٩ م.

مدونات لغويةٌ راصدةٌ Monitor

يُعد هذا النوع امتدادًا للمدونات اللغوية التاريخية؛ إذ إنها تراقب أو ترصد التطورات التي تطرأ على كلمةٍ أو مصطلحٍ ما بشكلٍ يوميٍ، أو أسبوعيٍ، أو شهريٍ، أو سنويٍ. ولذا يُطلق على هذا النوع أحيانًا اسم «المدونات اللغوية مفتوحة النهايات

open-ended». ومن أمثلة هذا النوع المدونة الإنجليزية الدولية الراصدة Global English Monitor Corpus، ومدونة لونجمان للغة الأمريكية المكتوبة Longman Written American Corpus. ويفيد هذا النوع كثيرًا في بناء المعاجم اللغوية؛ إذ إنها تمكن صانعي المعاجم من المتابعة الدقيقة لكل ما هو جديدٍ من كلماتٍ داخلةٍ على اللغة، أو تلك الكلمات التي تتغير معانيها عبر الوقت.

مدونات لغويةٌ تعليميةٌ Educational Corpora

تخدم فئةً معينةً من المستفيدين، وهم فئة دارسو ومُدَرِسو اللغات على السواء. وتضم نوعين من المدونات اللغوية:

مدونات الدارسين (للغات) Learner Corpora

عبارةٌ عن رصيدٍ من النصوص التي أعدها متعلمو اللغات. ولذا فهي تضم في العادة المقالات التي أنتجها هؤلاء المتعلمون. ويساعد هذا النوع من المدونات اللغوية في التحقق من الإنتاج اللغوي للطلاب من خلال المقارنة بين ما أنتجوه بعضهم البعض، وبين إنتاجهم اللغوي هذا وما ينتجه أبناء اللغة (وحينها لابد من توافر مدونةٍ لغويةٍ مقارِنةٍ تضم نصوصًا أنتجها أبناء اللغة).

ومن أبرز الأمثلة على هذا النوع المدونة اللغوية الدولية لدارسي الإنجليزية International Corpus of Learner Englsih (ICLE) التي تضم مقالاتٍ لمتعلمي الإنجليزية ذوي الأصول المختلفة (فرنسية، وسويدية، وألمانية، وغيرها). حيث تتم مقارنة هذه المقالات مقابل مدونةٍ لغويةٍ مقارِنةٍ لمقالاتٍ كتبها أبناء اللغة الإنجليزية، وهي مدونة لوفان لمقالات الإنجليزية الأصلية Louvain Corpus of Native English Essays LOCNESS.

ومن الممكن أيضًا استثمار هذا النوع من المدونات اللغوية في التحقق من الأخطاء اللغوية لدارسي اللغات؛ وذلك عن طريق وسم الأخطاء error tagging لما قام بإنتاجه الطلاب من مقالاتٍ. ومن ثم يمكن استرجاع وتقييم الأخطاء اللغوية التي يقع فيها دارسو لغةٍ ما على مختلف مستويات وأنواع هذه الأخطاء. الأمر الذي يُمَّكن المعلمين والقائمين على العملية التعليمية من التحقق من هذه الأخطاء وأسبابها، ومن ثم اتخاذ القرار المناسب بشأن معالجتها، وإمكانية تلافيها مستقبلًا.

ومثال على هذا النوع من المدونات اللغوية مدونة دارسي كامبريدج Cambridge Learner Corpus. ومن أبرز الأمثلة العربية على هذا النوع المدونة اللغوية لمتعلمي اللغة العربية Arabic Learner Corpus التي أعدتها جامعة ليدز.

مدونات لغويةٌ تربويةٌ أو تدريسيةٌ Pedagogical Corpora

تتكون من كل ما تَعرَّض له دارسو لغةٍ ما من نصوصٍ. أي أنها تضم كل الكتب، والمواد التعليمية، والتدريبات، والمواد السمعية والبصرية التي استخدمها هؤلاء الطلاب. وهكذا يتمكن الطلاب من استرجاع كافة السياقات المختلفة لأي كلمةٍ أو عبارةٍ تعلموها خلال فترة الدراسة. الأمر الذي يزيد من وعيهم وإدراكهم لما تعلموه. كما يمْكن استخدام هذا النوع أيضًا في المقارنة بين محتواه وبين ما تحويه مدونةٌ لغويةٌ عامةٌ أو مدونةٌ لغويةٌ لنصوصٍ مجمَّعةٍ من أبناء اللغة؛ وذلك بهدف التأكد من أن ما يتعرض له الطلاب يعكس فعليًا الواقع اللغوي المستخدم في المجتمع (Hunston, 2002, pp. 14-17; Blecha, 2012, pp. 11-22; MacMullen, 2003).

يُذكر أن هناك نوعًا أخذٌ في التطور يُطلق عليه ما وراء المدونات اللغوية meta corpora أو مدونة المدونات اللغوية corpus of corpora وهو عبارةٌ عن مدونةٍ لغويةٍ معتمِدةٍ على واجهة استخدام تُمكّن من البحث والاسترجاع في عددٍ من المدونات اللغوية بغرض إجراء تحليلٍ فيما بينها، أو لتصميم أدوات تحليلٍ معياريةٍ (MacMullen, 2003).

الاستخدام المنهجي للمدونات اللغوية

بعدما تم التأكد من أوجه الإفادة من المدونات اللغوية في مختلف المجالات، وكذلك مراحل الإنشاء، والأنواع، يطرح هذا السؤال نفسه: هل المدونات اللغوية منهجٌ أم نظريةٌ؟

غير إن الكثير من الباحثين في هذا التخصص ينظرون إلى المدونات اللغوية وفق الطريقة المنهجية التي يتم توظيفها في التحليل. فعلى سبيل المثال يرى البعض (Biber, Conrad & Reppen, 1998, p. 4) أن المدونات اللغوية هي مقاربةٌ منهجيةٌ approach تتسم بأربع خصائص رئيسةٍ، وهي أنها:

١. تجريبية أو واقعيةٌ empirical في تحليلها للنصوص القائمة على اللغة الطبيعية.
٢. تستخدم مجموعةً ضخمةً من النصوص كونها أساسًا لعملية التحليل.

٣. تستثمر الحاسب الآلي في التحليل.
٤. تعتمد على كلٍ من المنهج الكمي والمنهج النوعي في التحليل.

الخلاصة

ركز هذا الفصل على المدونات اللغوية كونها أداةً بحثيةً ومقاربةً منهجيةً. وفي هذا السياق تم تسليط الضوء على مجالات الإفادة منها في كلٍ من: علم اللغة، وتدريس وتعلم اللغات، وعلم اللغة الاجتماعي، وصناعة المعاجم، والترجمة، ودراسة التوجهات الفكرية (الأيديولوجيا)، وعلم المعلومات، وصناعة المكانز، والمعلوماتية الجنائية. كما ناقش الفصل مراحل إنشاء المدونات اللغوية وأساليب تطويرها، وأنواعها المختلفة (مدونات لغويةٌ اختباريةٌ، ومدونات لغويةٌ بحثيةٌ، ومدونات لغويةٌ تعليميةٌ)، وكيفية استثمارها منهجيًا.

الفصل الثالث

معالجة البيانات اعتمادًا على المدونات اللغوية

التمهيد
عمليات معالجة البيانات على مستوى الإدخال
ترميز المدونات اللغوية
تحشية المدونات اللغوية
أهمية تحشية المدونات اللغوية
النقد الموجه لتحشية المدونات اللغوية
طرق تحشية المدونات اللغوية
أنواع تحشية المدونات اللغوية
عمليات معالجة البيانات على مستوى الإخراج
تكشيف الكلمات
تكشيف النصوص
تكشيف الكلمات المفتاحية في السياق
المشكلات الفنية في تكشيف المدونات اللغوية
المشكلات اللغوية في تكشيف المدونات اللغوية
قوائم تردد الكلمات
توليد الكلمات المفتاحية
تحليل التجمعات العنقودية
الخلاصة

التمهيد

تحققنا في الأجزاء السابقة للكتاب هذا من المدونات اللغوية، وذلك في ضوء ماهيتها وأهميتها، ومميزاتها، ومواصفاتها، وما يمكن أن تقدمه وتجيب عنه، واستخداماتها في مجالات المعرفة المختلفة، ومراحل إنشائها، وأنواعها. ثم نأتي بهذا الفصل لنحاول فيه توضيح كيفية معالجة البيانات في المدونات اللغوية.

وبوجهٍ عامٍ، تتم معالجة البيانات اعتمادًا على المدونات اللغوية على مستويين؛ مستوى الإدخال، ومستوى الإخراج. وتشمل عمليات معالجة البيانات التي تتم على مستوى الإدخال كلًا من عمليتي ترميز المدونات اللغوية marking-up textual corpora، وتحشية المدونات اللغوية annotating textual corpora. بينما تشمل عمليات معالجة البيانات التي تتم على مستوى الإخراج كلًا من عمليات تكشيف الكلمات word indexing، ووضع الكلمات في قوائم word listing، وتوليد الكلمات المفتاحية keywords generation، وتحليل التجمعات العنقودية clustering.

عمليات معالجة البيانات على مستوى الإدخال

تُعد عمليات معالجة البيانات في المدونات اللغوية على مستوى الإدخال من أصعب العمليات الفنية في إعداد وتجهيز النصوص. ذلك لأنها تتطلب تدخلًا بشريًا في معظم مهامها. ولذا فهي مستهلِكة للوقت والجهد والتكلفة، غير إن المدونات اللغوية المُجهزة بمثل هذه العمليات الفنية تنتج عنها تحليلاتٍ ونتائج غاية في الأهمية إذا ما أُريد الوقوف بدقةٍ على بعض الظواهر اللغوية. وتتضمن هذه العمليات ترميز المدونات اللغوية، وتحشية المدونات اللغوية.

ترميز المدونات اللغوية

يُقصد به إضافة مجموعةٍ من حقول الميتاداتا أو الأكواد المعيارية إلى النصوص أو الوثائق المختزنة في المدونات اللغوية، بحيث تتوافر معلومات عن نصوص هذه الوثائق تتيح التحكم في أشكالها formats، وطرق معالجتها، مثل: إعطاء معلوماتٍ عن مصدر المدونة اللغوية، وتاريخ النص، ونوع النص، والمكان، وعدد المتحدثين، وتاريخ الميلاد، والمهنة، والمستوى التعليمي... إلخ. ويمكن وضع هذه الحقول وبياناتها في بداية كل ملفٍ رأسًا header، أو في ملفاتٍ مستقلةٍ مرتبطةٍ بالملف الأساس (Margaretha & Lüngen, 2014).

ولترميز المدونات اللغوية أهميةٌ خاصةٌ يمكن إيضاحها في مجموعة النقاط التالية:

١. أن بيانات المدونات اللغوية قائمةٌ بالأساس على رصيدٍ من النصوص الأصلية authentic؛ أي في ضوء الاستعمال الطبيعي للنصوص. إلا أنها بمجرد اختزانها في الملفات يكون قد تم إخراجها من سياقاتها الطبيعية؛ فيأتي دور الميتاداتا بما تقدمه من حقولٍ للوصف لتستعيد هذه السياقات المفقودة، وتُمكن القائمين على المدونات اللغوية من ربط النصوص بهذه السياقات مرةً أخرى. وكلما ازدادت حقول الوصف في الميتاداتا، ازدادت إمكانية تعويض الفقد السياقي للنصوص، وبالتالي ازدادت إمكانية إعادتها للاستعمال الطبيعي الذي ولدت فيه.

٢. على الرغم من أن حفظ النصوص يتم عادةً في ملفاتٍ، كلٌ بعنوانٍ يُمَيزه، إلا أن حقول الميتاداتا توفر معلوماتٍ إضافيةً عن هذه النصوص، تشمل بين ما تشمل: نوع النصوص المكتوبة، والعوامل والمتغيرات الاجتماعية واللغوية التي تحكم النصوص المنطوقة... إلخ.

٣. يتم اللجوء إلى ترميز المدونات اللغوية عندما يتم حذف الجداول وغيرها من الإيضاحيات من النصوص المكتوبة، فيُشار إليها بعلاماتٍ تعويضيةٍ تدل على أماكنها وأنواعها. بينما يتم ترميز النصوص المنطوقة للتعبير عن الملامح المساعِدة المميزة للحديث؛ كالضحك، والبكاء، والتثاؤب... إلخ.

٤. إضافة تعليقاتٍ خاصةٍ بالأعمال التحريرية التي تتم على المدونة اللغوية.

ومن أشهر المواصفات المستخدَمة في ترميز المدونات اللغوية إطارُ كاكاو COCOA، ومعيار دبلن كور Dublin Core، ومعيار تشفير المدونات اللغوية Corpus Encoding Standard CES، ومبادرة تشفير النصوص Text Encoding Initiative (TEI)، ويَستخدم الأخيران عادةً لغة الترميز المُعَمَمَة القياسية SGML ولغة الترميز القابلة للتمديد XML (Margaretha & Lüngen, 2014).

يُذكر أنه في ضوء ترميز المدونات اللغوية يمكن تقسيمها إلى نوعين: مدونات لغويةٌ خامٌ raw corpora، ومدونات لغويةٌ مرَمَّزةٌ marked-up corpora.

تحشية المدونات اللغوية

يُقصد بها إجراء تحليلٍ لغويٍ لنصوص المدونات اللغوية لاستخراج بعض التفسيرات والمعلومات اللغوية؛ مثل: وسم أقسام الكلم part-of-speech (POS) tagging، والإعراب التركيبي، وغيرهما...

وتعرف تحشية المدونات اللغوية Corpora Annotation بأنها مجموعةٌ من التحليلات والمعالجات اللغوية التي تتم على النصوص بهدف إضفاء توصيفٍ دقيقٍ عليها. ومن الممكن أن تتم تحشية المدونات اللغوية في عدة مستوياتٍ وبأشكالٍ مختلفةٍ: فعلى المستوى الصوتي، من الممكن أن تكون التحشية للمقاطع الصوتية (تحشيةٌ صوتيةٌ)، أو تحشيةٌ للخصائص العَروضية prosodic features (تحشيةٌ عَروضيةٌ)[1]. وعلى مستوى الصرف، من الممكن أن تتم التحشية للسوابق prefixes، واللواحق suffixes، والجذور roots، والجذوع stems (تحشيةٌ صرفيةٌ). وعلى مستوى المعجم، فتتم التحشية لأقسام الكلم parts-of-speech (وسم أقسام الكلم POS tagging)، أو للأسر اللغوية lemmas (تجريد الكلمات أو وضع الكلمات في أسرٍ لغويةٍ

١ - أي تحديد ما إذا كان للكلمة، أو الجملة، أو الفقرة دلالةٌ إيجابيةٌ أم دلالةٌ سلبيةٌ.

lemmatization)، أو للحقول الدلالية (تحشيةٌ دلاليةٌ). وعلى مستوى النحو، فتتم التحشية بالتحليل التركيبي syntactic analysis (التحليل الإعرابي parsing، وبنوك أشجار النصوص treebanks، وتقويس الكلمات bracketing). أما على مستوى الخطاب discourse، فتتم التحشية للعلاقات الإضمارية anaphoric (تحشية المصاحبة المرجعية coreference)، أو المعلومات البراجماتية أو التداولية (تحشيةٌ براجماتيةٌ)، أو الخصائص الأسلوبية، مثل عرض الكلام والمعتقدات (تحشيةٌ أسلوبيةٌ).

وتعد تحشية أقسام الكلم أشهر هذه الأنواع وأكثرها استعمالًا. كما أن التحليل التركيبي أخذٌ في التطور والتطبيق على نحوٍ متسارعٍ، بينما لا تَلقى أنواعٌ أخرى للتحشية الاهتمام المطلوب؛ كالتحشية الخطابية والتحشية البراجماتية (Abbas, 2012 .(Zaghouani & Dukes, 2014

أهمية تحشية المدونات اللغوية

تمثل التحشية قيمةً مضافةً للأسباب التالية:

١. تتيح إمكانية استرجاع واستخراج المعلومات وفقًا لنوع التحشية الذي تنتمي إليه كل كلمةٍ، فمثلًا من الصعوبة بمكانٍ أن نسترجع الصفات الموجودة على يسار الكلمات اعتمادًا على الشكل الإملائي أو السياق فقط، ولكن الأمر يتطلب التحديد المسبق بوسم «صفة»؛ فلولا هذا الوسم لكان من الممكن أن تكون النتيجة المسترجعة لمفعولٍ أو فاعلٍ أو غير ذلك من الأسماء.

٢. تفيد المحللين والباحثين في إجراء تحليلاتٍ إضافيةٍ على لغاتٍ يجهلونها (Leech, 1997, p. 2). فعلى سبيل المثال يمكن لباحثٍ يجهل اللغة العربية أن يجري بعض التحليلات على مدونةٍ لغويةٍ موَّسَمَةٍ tagged تمكنه من التحقق من أقسام الكلم بها دونما الحاجة إلى الإلمام باللغة العربية إلمامًا كاملًا.

٣. ربما تؤدي تحشية المدونات اللغوية وظائف أخرى غير التي أُنشأت من أجلها؛ بحسب الاحتياجات الآنية للباحثين.

النقد الموجه لتحشية المدونات اللغوية

على الرغم مما تمثله تحشية المدونات اللغوية من أهميةٍ، إلا أن هناك بعض الانتقادات التي وجهت لها، منها:

١. أن تحشية المدونات اللغوية تؤدي إلى نوعٍ من التشويش على النصوص المسترجَعة؛ ولذا يجب أن تكون هناك طريقةٌ أخرى تُعرَض بمقتضاها النصوص في شكلها البسيط plain text، بحيث تكون التحشية على هامش النصوص، أو في شكل إشاراتٍ جانبيةٍ labels. وهذا بالفعل ما التزمت به بعض برمجيات معالجة المدونات اللغوية في الآونة الأخيرة، مثل: برنامج وورد سميث WordSmith، وبرنامج مونو كونك MonoConc، وبرنامج سارا SARA، وبرنامج زايرا Xaira، وغيرها، والتي تتيح إمكانية إخفاء التحشية عن نتائج البحث المسترجعة بحيث تُتاح النصوص البسيطة فقط (Hunston, 2002, p. 94).

٢. أن التحشية قد تَفرِض على المستفيدين من المدونات اللغوية تحليلًا لغويًا يمثل عبئًا عليهم. فعلى الرغم من أن تحشية المدونات اللغوية تتيح قدرًا من التفسير للنصوص، إلا أن ذلك ليس بالضرورة أن يَلقى قبول كل المستفيدين. ولذا وجب أن يتمكن المستفيدون من إظهارها أو إخفائها وقتما أرادوا (McEnery, 2003, p. 456).

٣. قد تتحول التحشية من كونها قيمةً مضافةً إلى كونها قيمةً زائدةً overvalue تقلل من فرص إتاحة، وتحديث، وتمدد المدونات اللغوية؛ نظرًا لما تتطلبه عملية التحشية من جهدٍ ووقتٍ وتكلفةٍ (Hunston, 2002, pp. 92-93).

٤. قد لا تكون التحشية بالدقة والاتساق المطلوبين. وأيًا كانت الطريقة التي تتم بها التحشية، فإنه لا يمكن بحالٍ أن تؤدي إلى نتائج دقيقةٍ ومتسقةٍ مئةٌ بالمئة (Hunston, 2002, p. 91). ولذا يَنصح الكثيرُ بتدخل الاختصاصيين عقب التحشية الآلية أو شبه الآلية بالمراجعة والتدقيق للحد قدر الإمكان من نسبة الأخطاء.

ومن الجدير بالذكر أن تحشية المدونات اللغوية جعلت البعض يقسمها إلى نوعين: مدونات لغويةٌ مُحَشَّوَّةٌ annotated، ومدونات لغويةٌ غير مُحَشُّوَّةٍ أو خامٌ.

طرق تحشية المدونات اللغوية

تتم تحشية المدونات اللغوية بطرقٍ ثلاث: الطريقة الآلية، والطريقة شبه الآلية، والطريقة اليدوية. ففي الطريقة الأولى تقوم البرمجيات بالتحشية بناءً على قواعد وخوارزميات تم إعدادها مسبقًا بواسطة مبرمجين. وعلى الرغم من أن هذه الطريقة مكلفةٌ ومستنفدةٌ للوقت، إلا أنه بمجرد الانتهاء من هذه الأداة البرمجية فإنه يمكن تحشية كم ضخمٍ من النصوص بالسرعة والاتساق المطلوبين. وهذه الطريقة تعمل بكفاءةٍ عاليةٍ مع بعض اللغات؛ كاللغة الإنجليزية، والفرنسية، والأسبانية بمعدل خطأ قد لا يتجاوز ٣٪. وهنا قد يتدخل البشر في إجراء التصويبات المطلوبة، حيث توفر بعض برمجيات التحشية إمكانية التدخل البشري لحل حالات الخطأ أو اللبس التي تستعصي على البرامج الآلية. وينتج عن هذه الطريقة شبه الآلية نتائج مسترجعةٌ أكثر موثوقيةً من نظيرتها في الطريقة الآلية كليةً.

أما التحشية اليدوية فتتم بالاعتماد كليةً على محللٍ بشريٍ دون أي أداةٍ برمجيةٍ. ونظرًا إلى أن هذه الطريقة مكلفةٌ ومستنفدةٌ للوقت، فإنه يُفضَّل استخدامها مع المدونات اللغوية الصغيرة (Hovy & Lavid, 2010, pp. 13-14; McEnery, Xiao & Tono, 2006, p. 33).

أنواع تحشية المدونات اللغوية

هناك أنواعٌ عدةٌ لتحشية المدونات اللغوية، منها:

وسم أقسام الكلم parts-of-speech (POS) tagging

يُطلق عليه أيضًا الوسم القواعدي grammatical tagging، أو الوسم الصرفي النحوي morpho-syntactic tagging. ويتم فيه تحديد القسم الذي تنتمي إليه كل كلمةٍ (اسمٌ، وفعلٌ، وحرفٌ). ويُعد وسم أقسام الكلم أول نوعٍ استُخدم في تحشية المدونات اللغوية، والأكثر شهرةً بين الأنواع الأخرى. كما أنه يُشكّل الأساس الذي تتم من خلاله أشكال وأنواع التحليل الأخرى. وتبدأ أوجه الإفادة من وسم أقسام الكلم بإزالة اللبس disambiguation عن الكلمات المتجانسة، وحتى احتساب فئات الكلمات داخل النصوص. وتعتمد الكثير من تحليلات المدونات اللغوية على هذا النوع من التحشية، مثل تحليل التلازم اللغوي[1] collocation (Hunston, 2002).

١ – يُسميه البعض أيضًا: تلازم المصطلحات، أو التلازم المعجمي، أو المصاحبة اللغوية، أو التلازم اللفظي. وهي كلماتٌ تتصاحب أو تتلازم أو تَقترن دائمًا بكلماتٍ أخرى، سواءٌ لسببٍ، أو دون سببٍ ظاهرٍ أو منطقيٍ.

ومع بلوغ وسم أقسام الكلم مراحل متطورةً آليًا، فإنه يمكن أن يصل معدل التحقيق في النتائج المسترجعة إلى مستوياتٍ عاليةٍ تعتمد عليها الأبحاث والدراسات. وتعد أداة وسم أقسام الكلم المعروفة بنظام كلوز CLAWS من أشهر الأدوات المعالِجة للغة الإنجليزية التي قامت عليها جامعة لانكستر (Garside, Leech & Sampson, 1987).

وتزيد بعض الخصائص المميزة للغة العربية من أهمية هذا النوع من أنواع تحشية المدونات اللغوية. فظاهرة الجناس، والحروف غير المشَّكَّلَة، والسوابق واللواحق في اللغة العربية، على سبيل المثال، تؤدي إلى حدوث الكثير من المشكلات أثناء عملية التحليل. ولكن مع وسم أقســـام الكلـم يمكن التغلب على مثل هذه المشكلات (Van Mol, 2000, pp. 1-4).

ويلاحظ أن التقسيم العربي التقليدي يحتاج إلى مراجعةٍ وتفصيلٍ أكثر عند وسم أقسام الكلمات، حتى يكون التحليل أكثر فائدةً ودقةً (Buckwalter & Parkinson, 2011).

وجديرٌ بالذكر أن كتب النحو العربي تزخر بمصطلحاتٍ تتعلق بتصنيف الكلمات العربية ويمكننا الاستفادة منها، مثل: ضمير، واسم موصول، واسم مشتق.... وفعل لازم، وفعل متعد... وحرف جر، وحرف نصب، وحرف جزم... ويجب وسم الكلمات بمثل هذه التصنيفات لإجراء كثير من الدراسات والأبحاث (Sawalha & Atwell). ومن أشهر أدوات وسم أقسام الكلمات العربية برنامج بَكوالتر Buckwalter.

ويتم وسم أقسام الكلم باستخدام أشكالٍ مختلفةٍ من التشفير؛ فمن الممكن أن تُربط الوسيمة tag بالكلمة عن طريق شرطةٍ سفليةٍ underscore، أو علامة &، أو علامة = . كما في النص التالي المأخوذ من صحيفة الجزيرة (Khoja, Garside & Konwles, 2001):

بعث_VPSg3M خادم_NCSgMNI الحرمين_NCDuMAD الشريفين_NCDuMGD الملك_NCSgMND فهد_NP بن_
NCSgMNI عبد_NCSgMAI العزيز_NCSgMAD آل_R سعود_NP برقية_NCSgFNI تهنئة_NCSgFGI الى_PPr
فخامة_NCSgFGI الرئيس_NCSgMGD الكسندر_RF أواسنيفيسكي_RF رئيس_NCSgMNI جمهورية_NCSgFGI
بولندا_RF بمناسبة_PPr_NCSgFGI اليوم_NCSgMAD الوطني_NCSgMND لبلاده_PPr_NCPlFGI_NPrPSg3M
وأعرب_PC_VPSg3M الملك_NCSgMND المفدى_NCSgMAD باسمه_PPr_NCSgMGI_NPrPSg3M وباسم_
PC_PPr_NCSgM شعب_NCSgMGI وحكومة_PC_NCSgFGI المملكة_NCSgFGD العربية_NCSgFGD السعودية_
NCSgFGD عن_PPr اخلص_NCSgFNI التهاني_NCPlMND متمنياً_NCSgMAI لفخامته_
PPr_NCSgFGI_NPrPSg3M دوام_NCSgMNI الصحة_NCSgFGD والسعادة_PC_NCSgFGD ولشعب_
PC_PPr_NCSgMGI بولندا_RF الصديق_NCSgMND الازدهار_NCSgMND الدائم_NCSgMN ._PU

وتشير أول وسيمة (VPSg3M) للفعل «بعث» في هذا النص على سبيل المثال إلى الآتي:

VP	Verbal Phrase	جملة فعلية
Sg	Singular	مفرد
3	Third Person	غائب
M	Masculine	مذكر

كما يمكن لهذه الكلمة أن يتم وسمها باستخدام لغة الترميز المُعَمَمَة القياسية SGML، أو باستخدام لغة الترميز القابلة للتمديد XML كالتالي:

<w POS=VPSg3M>بعث</w>

وتُفضل بعض برمجيات المدونات اللغوية وسم أقسام الكلم باستخدام الشرطة السفلية فقط، مثل وورد سميث WordSmith، ومونوكونك MonoConc. بينما تُفضل برمجياتٌ أخرى مثل: سارا SARA، وزيرا Xaira استخدام لغة الترميز المُعَمَمَة القياسية SGML أو لغة الترميز القابلة للتمديد XML.

ومن الأمور المهمة التي يتعين وضعها في الحسبان عند وسم أقسام الكلم هو كيفية تجزيء النصوص إلى هياكل من الكلمات (tokenization) word tokens. ففي اللغة الإنجليزية تتم تجزئة هياكل الكلمات في النصوص المكتوبة عن طريق تحديدها delimit بمسافةٍ قبلها وبعدها. ولكن في اللغة العربية نجد أن الكثير من كلماتها متصلة clitic بضمائر، أو أدوات عطف، أو حروف جر دون وجود مسافات بينها. ناهينا عن الطبيعة الاشتقاقية العالية التي تتميز بها العربية. مما يزيد الأمر تعقيدًا إذا ما أُريد تحشية أقسام الكلمات العربية. ففي برنامج بَكوالتر تتم تجزئة جملة مثل «وسيكتبونها» على النحو التالي:

[CONJ +
FUTURE PARTICLE +
IMPERFECT VERB PREFIX +
IMPERFECT VERB +
IMPERFECT VERB SUFFIX MASCULINE PLURAL
3RD PERSON +
OBJECT PRONOUN FEMININE SINGULAR]

فهذه الجملة المتصلة تتكون من حرف عطفٍ (و)، وأداةٍ للمستقبل (سـ)، وسابقةٍ

prefix صرفيةٍ (يـ)، وجذعٍ stem (كتب)، ولاحقةٍ suffix صرفيةٍ (ون)، ومفعولٍ به وقع ضميرًا (ها) (Mohamed & Kubler, 2010; Habash & Rambow, 2005).

تجريد الكلمات lemmatization

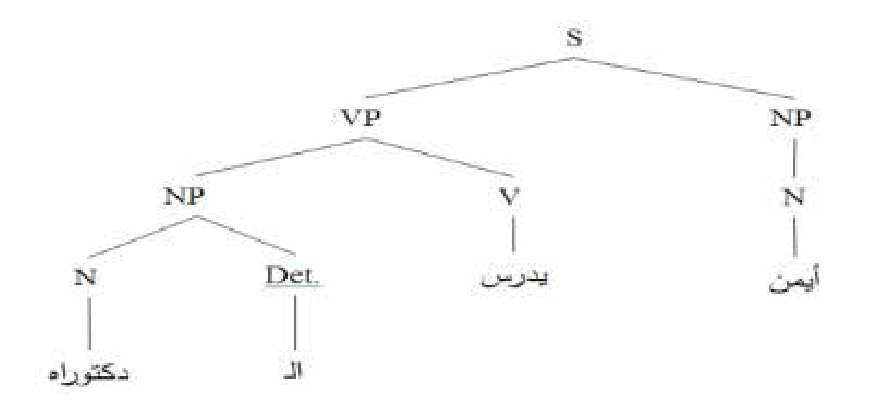
هو أحد أنواع تحشية النصوص الذي يقوم باختزال أو إعادة الصور المتصرفة المختلفة للكلمة الواحدة إلى صورتها المعجمية الأساس كما تظهر كونها مداخل معجميةً. ويتم ذلك من خلال حذف الزوائد. فالأسماء يتم ردها إلى المفرد النكرة (المذكر إن أمكن)، مثل إرجاع: مشروعان، ومشروعات إلى «مشروع». والأفعال يتم إرجاعها إلى الماضي الغائب المفرد المذكر، مثل إعادة: يكتبان، ويكتبون، واكتب... إلى الجذع «كتب». ومن هنا نجد أن تجريد الكلمات له أهميةٌ كبيرةٌ في الدراسات المعجمية وصناعة المعاجم بوجهٍ عامٍ (El-Shishtawy & El-Ghannam, 2012).

وتزداد الإفادة من تجريد الكلمات كلما ازدادت قابلية اللغة للاشتقاق. واللغة العربية لغةٌ عالية الاشتقاق؛ ولذا يؤدي تجريد الكلمات دورًا بارزًا في معالجة نصوصها الطبيعية. أما لغةٌ أخرى مثل الصينية فهي غير قابلةٍ للاشتقاق؛ ولذا ليست هناك حاجةٌ للإفادة من تجريد كلماتها. بينما لغةٌ كاللغة الإنجليزية فتتميز ببساطة نظامها الصرفي؛ ولذا ليست هناك صعوبةٌ في تجريد كلماتها، ولذا أيضًا قلما نجد في المدونات اللغوية الإنجليزية إمكانية معالجتها من خلال تجريد كلماتها (Leech, 1997).

التحليل الإعرابي parsing

بمجرد وسم أقسام الكلم في المدونات اللغوية فإنه يمكن وضع هذه الفئات الصرفية النحوية في مستوىً أعلى من العلاقات النحوية. أي تحليل الجمل إلى مكوناتها الأساس constituents. وقد يتم وضع هذه المكونات الأساس بين أقواسٍ؛ ولذا يطلق على التحليل الإعرابي أحيانًا التقويس bracketing. ويتم عادةً التحليل الإعرابي للجمل في شكلٍ تفريعيٍ يشبه الشجرة يطلق عليه: أشجار تركيب العبارة phrase structure trees. وتتجمع هذه الأشجار مع بعضها البعض لتمثل الجمل والعبارات المكوِنة للمدونة اللغوية، وتسمى حينئذٍ: بنوك أشجار النصوص treebanks. فمن الممكن تحليل جملة «أيمن يدرس الدكتوراه» على النحو التالي:[١]

١- s=sentence أي جملة. NP=Noun Phrase أي عبارة اسمية. VP=Verbal Phrase أي عبارة فعلية. Det.=Determiner أي المحدد أو الأداة.



ومن أشهر بنوك أشجار النصوص العربية هو البنك الشجري العربي من جامعة بنسلفانيا (Maamouri, Bies, Jin, & Buckwalter, 2003) يليه بنك براغ الشجري (Jan, et al., 2004). وذلك الذي تم إعداده في جامعة ليدز البريطانية (Dukes, Atwell & Sharif, 2010). وذلك الذي تم إنشاؤه في جامعة كولومبيا الأمريكية (Habash & Roth, 2010).

وحريٌّ بالذكر أن التحليل الإعرابي عمليةٌ فنيةٌ أكثر تفصيلًا من وسم أقسام الكلم؛ لأن الأخير يمثل عمليةً فرعيةً من الأساس (Leech & Wilson, 1996). وعلى الرغم من ذلك فإن معدل التحقيق في النتائج المسترجعة من الأخير يكون عادةً أعلى منه في الأول (Barbu, Evans & Mitkov, 2002)؛ ولذلك يتم عادةً تصويب أخطاء المدونات اللغوية المُعْرَبة يدويًا من أجل الارتفاع بمستوى التحقيق، كما هو الحال في بنك أشجار نصوص بِن Penn Treebank الذي أنشأه قسم الحاسب الآلي وعلم المعلومات التابع لجامعة بنسلفانيا University of Pennsylvania.

والتحليل الإعرابي إما أن يكون هيكليًا skeleton أو تفصيليًا detailed؛ فالأول يقدم تحليلًا سطحيًا shallow يقف عند حدود المكونات الأساس للجملة، كما في المثال السابق، أما الثاني فيقدم تحليلًا أكثر تفصيلًا ينزل بمستوى التحليل للعبارة الاسمية، على سبيل المثال، ليبين العدد، والنوع... وغيرهما (Swift, Allen & Gildea, 2004).

التحشية الدلالية semantic annotation

تسمى أحيانًا وسم معاني الكلمات word sense tagging. ويتم في هذا النوع من التحشية تخصيصُ مجموعةٍ من الأكواد التي تعكس الخصائص والحقول الدلالية لكلمات النصوص. وهذا النوع من الأهمية بمكانٍ في دراسات تحليل المحتوى. ويتميز بأنه أكثر صعوبةً من وسم أقسام الكلم، والتحليل الإعرابي والنحوي؛ لأنه يعتمد بشكلٍ رئيسٍ على المعرفة knowledge-based، ويتطلب أُنطولوجيات، ومعاجم، ومكانز معدةً سلفًا كي يؤدي مهامه.

وثمة نوعان على الأقل من التحشية الدلالية؛ يحدد النوع الأول العلاقات الدلالية بين المكونات الأساس للجمل داخل النصوص، كما هو في بنك أشجار نصوص بِن السابق ذكره. ويُطلق على هذا النوع أيضًا التحليل الإعرابي الدلالي semantic parsing، ويعتبره البعض نوعًا من التحشية النحوية. بينما يحدد النوع الثاني، وهو الأشهر استخدامًا، الخصائصَ الدلالية لكلمات النصوص (Thompson & Mooney, 2003).

تحشية المصاحبة المرجعية coreference أو العائدية anaphoric

يعتبرها البعض أحد أنواع التحشية على مستوى الخطاب. ويطلق على المدونات اللغوية ذات هذا النوع من التحشية اسم المدونات اللغوية المَحْشُوَّة annotated، والتي تتسم بمحدودية انتشارها، على الرغم من أهميتها البالغة في تعيين علاقات الإضمار والمصاحبة المرجعية بين النصوص. وربما لا توجد غير أداةٍ وحيدةٍ لمعالجة النصوص العربية في هذا الشأن، وهي الأداة التي طورها مايتكوفMitkov ، وبيلجويث Belguith، وماجورزاتا Malgorzata (1998).

وتهتم هذه التحشية بتحديد العلاقة بين كيانين في النصوص؛ أحدهما إحالةٌ لغويةٌ قبليةٌ anaphor يشير إلى كيانٍ آخر، يسمى المُحال إليه referent، مذكورٍ قبله antecedent أو قد يكون إحالةً لغويةً بعديةً cataphor يشير إلى كيانٍ مذكورٍ بعده postcedent، بما يكفل التماسك والترابط بين النصوص من خلال اعتمادها على التكرار، أو الحذف، أو الإبدال أو التعويض.... ويتضمن ذلك:

١. الإحالة اللغوية الضميرية: تشمل ضمائر الغائب، وأسماء الإشارة، والأسماء الموصولة.
٢. الإحالة المعجمية: تحدث عندما يكون المُحال إليه صفةً مُعَرَّفَةً أو اسم علمٍ؛ وتؤدي وظيفةً معينةً، مثل: الترادف، أو التعميم، أو التخصيص...

٣. الإحالة اللغوية المقارِنة: تشمل: المحددات المعجمية، مثل: آخر، وأخرى، وواحد، كأن نقول: أعدت بحثين؛ واحدٌ في علم اللغة، والآخر في علم المعلومات. وتشمل أيضًا أسماء التفضيل (الشاذلي، ١٩٩٨).

والمثال التالي لإحدى الإطارات schemes المستخدمة في تحشية المدونات اللغوية وفقًا للمصاحبة المرجعية (Hammami, Belguith & Hamadou, 2009):

```
- <p>
  - <s>
      كنا فعلت
      <exp id="e2" cat="Np" fc="sujet">خديجة</exp>
    </s>
  - <s>
      حملت على عاتق :
    - <exp id="e3" cat="pln" dist="1" rec="true">
        <ptr type="coref" src="e2" />
        ها
      </exp>
      أشياء كثيرة,
    </s>
  - <s>
```

التحشية البراجماتية أو التداولية pragmatic annotation

تُعتبر أيضًا أحد أنواع التحشية على مستوى الخطاب. وهي تهتم أكثر ما تهتم بتحليل المرجعية reference، وتفسير وتوليد الأفعال الكلامية speech acts، وعلاقات التركيب والترابط الخطابي، والافتراض (Bunt & Black, 2000). وقليلةٌ هي المدونات اللغوية المَحْشُوَّة تداوليًا، ولاسيما العربية منها. وقد حاولت دعاء وآنا (Samy & González-Ledesma, 2008) تقديم إطارٍ قائمٍ على الـ XML لتحشية مدونةٍ لغويةٍ ثلاثية اللغة، ومنها اللغة العربية، على مستوى علامات الخطاب. ويقدم المثال التالى نموذجًا لتحشية إحدى علامات الخطاب في هذه المدونة:

```
و هي  -->  CONJUNCTION+PRONOUN
```

التحشية الأسلوبية stylistic annotation

يرتبط هذا النوع من تحشية المدونات اللغوية بالخصائص الأسلوبية للنصوص الأدبية (Leech, McEnery & Wynne, 1997)، من خلال ما يُعْرَف بتمثيل الكلام والأفكار، وطول الكلمات، ومدى إثراء المفردات، والخصائص المعجمية للكلمات

والتمثيلات (Abbasi, Chen & Salem, 2008). حيث يتم وضع النصوص في فئاتٍ؛ فئة النصوص المباشرة، وفئة النصوص المباشرة الحرة، وفئة النصوص غير المباشرة، وفئة النصوص الحرة غير المباشرة، وفئة تمثيل النصوص والكلام والأفكار، وفئة تمثيل الصوت والحالة الداخلية للنصوص، وفئة النصوص التقريرية report. وتفيد التحشية الأسلوبية كثيرًا في التحقق من النصوص مجهولة التأليف (McIntyre & et al., 2004). ونظرًا لصعوبة تطبيقها على النصوص بطريقةٍ آليةٍ، فإن التحشية الأسلوبية تتم عادةً بطريقةٍ يدويةٍ.

وسم الأخطاء error tagging

يرتبط هذا النوع من التحشية بالمدونات اللغوية لدارسي اللغات وتدريس اللغات بوجهٍ عامٍ. حيث يتم استخدام أكواد معينةٍ توضح نوع الخطأ اللغوي في النصوص. الأمر الذي من شأنه أن يسهم في الكشف عن الأخطاء اللغوية التي يقع فيها دارسو اللغات باستمرار في المستويات المختلفة. وتختلف إطارات وسم الأخطاء من مدونةٍ لغويةٍ لأخرى وفقًا لعدد وأنواع أكواد الأخطاء المستخدَمة. وبوجهٍ عامٍ، فإن أشهر أنواع الأخطاء: الحذف، والإضافة، وسوء الصياغة. فعلى سبيل المثال تَستخدم مدونة دارسي كامبريدج Cambridge Learner Corpus الكود F للدلالة على خطأ في شكل الكلمة، والكود M للدلالة على فقدان كلمةٍ أو حرفٍ... والكود R للدلالة على استبدال كلمةٍ أو جملةٍ مكان أخرى، والكود U للدلالة على كلمةٍ أو جملةٍ غير ضروريةٍ، والكود D للدلالة على خطأٍ في تصريف الكلمة (Nicholls, 2003).

أما في اللغة العربية فيمكن تصنيف الأخطاء اللغوية إلى: أخطاءٍ صوتيةٍ، وأخطاءٍ صرفيةٍ، وأخطاءٍ نحويةٍ، وأخطاءٍ معجميةٍ، وأخطاءٍ إملائيةٍ، وأخطاءٍ دلاليةٍ (العصيلي، ١٩٨٥؛ العتيق، ١٩٩٢). وقد استخدمت إحدى الدراسات الكود « إ » للدلالة على خطأٍ إملائيٍ، والكود «ص» للدلالة على خطأٍ صرفيٍ، والكود «ن» للدلالة على خطأٍ نحويٍ، والكود «س» للدلالة على خطأٍ أسلوبيٍ، والكود «ت» للدلالة على خطأٍ في علامات الترقيم (Alfaifi, Atwell & Abuhakema, 2013).

التحشية الموجهة نحو المشكلات problem-oriented annotation

وصفها هان (Haan, 1984) بأنها واقعةٌ بحثيةٌ يستخدم فيها المستفيدون المدونات اللغوية، المَحْشُوَّة أو غير المَحْشُوَّة، كي يضيفوا التحشية الخاصة بهم على النحو الذي يجيب عن أسئلة الدراسة. ولذا تُسمى أيضًا التحشية محددة البحث research specific. ويختلف هذا النوع من التحشية عن الأنواع الأخرى السابق ذكرها من ناحيتين:

١. أنها سهلةٌ وموفرةٌ للوقت؛ فليس من الضروري أن تتم تحشية كل كلمةٍ أو جملةٍ، بل يتم فقط تحشية ما له علاقةٌ بأهداف البحث والدراسة. ولذلك يعتبر البعض التحشية العائدية ضمن هذا النوع.

٢. أنها تتيح للمستفيدين حرية اختيار إطارات التحشية التي تتناسب وأهداف دراساتهم؛ أي أنها لا تلزمهم بإطارٍ معينٍ للتطبيق.

التحشية المتضمنة embedded والتحشية القائمة بذاتها stand alone

تتم معظم أنواع تحشية المدونات اللغوية السابقة وهي مختلطةٌ بالوثائق الأصلية للمدونة اللغوية (تحشيةٌ متضمنةٌ). غير إن معايير ترميز المدونات اللغوية توصي دائمًا باستخدام تحشيةٍ منفصلةٍ في شكل إطاراتٍ لترميز النصوص، التي من أشهرها كما أسلفنا الذكر لغة الترميز المُعَمَمَة القياسية SGML ولغة الترميز القابلة للتمديد XML. حيث يتم ربط لغات البرمجة هذه بالوثائق الأصلية وغيرها من النصوص المَحْشُوَّة في شكل نصوصٍ فائقةٍ hypertexts (تحشيةٌ قائمةٌ بذاتها) (Thompson & McKelvie, 1997).

وتعالج التحشية القائمة بذاتها أحد عيوب التحشية بوجهٍ عامٍ، وهي التشويش على النصوص المسترجعة؛ إذ إنها بذلك تتيح إمكانية عرض نصوص الوثائق الأصلية في معزلٍ عن أي تشويشٍ يكتنفها بسبب التحشية. ومن المزايا الأخرى للتحشية القائمة بذاتها:

١. تقديمها في معزلٍ عن الوثائق الأصلية يجعلها تنأى بنفسها عن أي قيودٍ قانونيةٍ.

٢. تتم بسهولةٍ بصرف النظر عن شكل الوثائق الأصلية، كالوثائق التي تُسْمَح بقراءتها فقط read-only.

٣. تتجنب إنشاء وثائق غير مجديةٍ عمليًا.

٤. تتيح قدرًا من التداخل المتعدد بين التفريعات الهرمية لتمثيل النصوص.

٥. تقبل أكثر من نوعٍ واحدٍ من التحشية في آنٍ واحدٍ.

٦. تتيح تعديل أو إضافة مستوياتٍ أخرى من التحشية دون إحداث أي مشكلةٍ للتحشيات الموجودة بالفعل أو لأدوات وبرمجيات البحث (Gries & Berez, 2013).

عمليات معالجة البيانات على مستوى الإخراج

تتمثل في العمليات الفنية التي تُقدِم المخرجات للمستفيدين من خلال واجهات الاستخدام. وهذه العمليات الفنية تعتمد بالأساس على برمجياتٍ يتم تشغيلها على متون المدونات اللغوية. ولذا تتجسد الصعوبة هنا في تطوير هذه البرمجيات فقط. وبمجرد إيعاز هذه البرمجيات فإن عملها لا يأخذ بضع ثوانٍ في أغلب الأحيان.

تكشيف الكلمات

قبل الحديث عن تكشيف الكلمات يتعين أولاً توضيح المقصود بالكلمات لغويًا وآليًا. فالكلمات جمع «كلمة»، وتعني من الناحية اللغوية لفظةً واحدةً أو مجموعة ألفاظٍ دالةٍ على معنىً (عمر، ٢٠٠٨، في معجم اللغة العربية المعاصرة، ج ٤، ص ١٩٥٤). وتعني من الناحية الآلية وحدة المعالجة الرئيسة ضمن مجموعةٍ من النصوص (Encyclopedia of Information Science and Technology, Third Edition, p. 1824). والتعريف الأخير مرتبطٌ بشكلٍ مباشرٍ بـ «الكلمة» في نظر المدونات اللغوية التي تتعامل معها على أساس أنها هياكل كلماتٍ tokens؛ أي سلسلةٌ من الحروف أو التمثيلات التي يسبقها فراغٌ ويتبعها فراغٌ.

ويُطلق على تكشيف الكلمات أيضًا: التكشيف الاشتقاقي، أو التكشيف بالاقتطاف. وتعتمد المدونات اللغوية بشكلٍ أساسٍ على تكشيف جميع الكلمات الواردة بالنصوص كونها مداخل كشفيةً (تكشيفٌ حُرٌ أو تكشيف اللغة الطبيعية)، وقلما يُمارَس التقنين لأشكال المداخل أو التحكم فيها. وبوجهٍ عامٍ فإن الكلمات قد تُشتق أو تُقتطَف من النصوص الكاملة للمدونات اللغوية، أو من العناوين. وهذا يعني أن ثمة نمطين أو نوعين من تكشيف المدونات اللغوية، هما: تكشيف النصوص concordancing، وتكشيف الكلمات المفتاحية في السياق KeyWord In Context (KWIC) (عبد الهادي و زايد، ٢٠٠٠، ص ٣٩).

تكشيف النصوص

أي عمل كشافٍ هجائيٍ أو معجمٍ مفهرسٍ أو كشاف ألفاظٍ (concordance) للكلمات الواردة في الوثائق في سياقٍ محددٍ، دون تمييزٍ بين الكلمات؛ فكل كلمةٍ يشتمل عليها النص تعتبر مَدخلًا كشفيًا (عبد الهادي، ١٩٨٢، ص ص ٦٨-٦٩). ويُعد هذا النمط من أقدم أنماط التكشيف، إن لم يكن أقدمها على الإطلاق. وتعني كلمة concordancing الانسجام أو الاتفاق أو التلازم، غير أنها اكتسبت معنىً اصطلاحيًا يدل على نوعيةٍ متميزةٍ من التكشيف (قاسم، ٢٠٠٠، ص ص ١٩-٢٠). ويهتم تكشيف النصوص بمتن الإنتاج الفكري؛ حيث يتم تكشيف النصوص في استرجاع كل كلمةٍ ترد في نصٍ معينٍ، أو في مجموعة الأعمال الكاملة لأديبٍ أو شاعرٍ أو مؤلفٍ معينٍ، أو في تخصصٍ معينٍ، أو في مجموعةٍ من التخصصات (كالنصوص المقدسة، والقوانين، والدساتير... إلخ) كما وردت في سياقاتها المختلفة (عبد الهادي، ١٩٨٢، ص ٧١). وقد يكون هذا السياق فقرةً أو جملةً أو عبارةً أو سطرًا أو مَقْطَعًا؛ حيث تقع الكلمة (عبدالهادي و زايد، ٢٠٠٠، ص ص ٣٧-٣٩).

وقد يجعل تكشيف النصوص المخرجات أضخم من النصوص الأصلية؛ لأن الكلمات يتم ترتيبها هجائيًا في مداخل أو أسطرٍ توازي عدد الكلمات الرئيسة الواردة في النصوص، وليس بحسب عدد جمل السياق، التي تقع فيها الكلمات. فمثلًا إذا كانت لدينا جملةٌ مسترجعةٌ مكونةٌ من عشر كلماتٍ، فإن إخضاعها للتكشيف بالمدونة اللغوية يُخرجها في عشرة أمثال حجم النص الأصلي، وخصوصًا إذا مَثلت كل كلمةٍ بهذه الجملة اهتمامًا خاصًا. فعلى سبيل المثال قد ينصب اهتمام المتخصصين في علم المعلومات، أو القانون، أو التعليم... على الكلمات الرئيسة دون الكلمات المستثناة stop words التي يحصرونها عادةً في قوائم مستقلةٍ، إلا أن الكلمات المستثناة هذه قد تجتذب اهتمام المتخصصين في علم اللغة، أو الترجمة، أو صناعة المعاجم... ولذا نجد أن تكشيف نصوص المدونات اللغوية يشتمل دائمًا على جميع كلمات النص دون استثناء (عبدالهادي و زايد، ٢٠٠٠، ص ص ٣٧-٣٩).

ونظرًا لضخامة مخرجات عملية تكشيف النصوص، فإن برامج معالجة المدونات اللغوية لا تقدمها دفعةً واحدةً؛ بل إنها توفرها وفق كل كلمةٍ على حدةٍ.

وتُتاح الآن بعض برمجيات تكشيف النصوص concordancers مجانًا، مثل إحدى النسخ القديمة لبرنامج أدوات وورد سميث WordSmith Tools الذي يقوم عليه مايك

سكوت Mike Scott الباحث بجامعة ليفربول University of Liverpool وبالتعاون مع جامعة أُكسفورد Oxford University (Scott, 2016). وبرنامج «أدوات معالجة المدونات اللغوية العربية» الذي يقوم عليه عبد المحسن الثبيتي وزملاؤه بمدينة الملك عبد العزيز للعلوم والتقنية (Almujaiwel & Al-Thubaity, 2016).

تكشيف الكلمات المفتاحية في السياق

الكلمات المفتاحية هي عناصر لغويةٌ تحمل أبرز المعاني التي تتضمنها أي وثيقةٍ. والكلمات المفتاحية أو الدالة وفقًا لهذا المفهوم عبارةٌ عن كلماتٍ تشير إلى موضوعٍ من الموضوعات التي تعالجها الوثيقة، وتُستخدم نقاطَ إتاحةٍ، وبالتالي تُستخدم مداخلَ تعكس المحتوى الموضوعي للوثائق.

وقد ابتكر هذا النوع من التكشيف هانز بيتر لون. ويسمى أحيانًا تكشيف العناوين؛ إذ يقوم هذا النوع من التكشيف بالترتيب الهجائي للكلمات المفتاحية الواردة بالعناوين (عبد الهادي، ١٩٨٢، ص ٧٢) التي تم جمعها بالمدونات اللغوية، مع بيان سياقاتها التي وردت بها؛ وذلك بتسجيل بقية العنوان. وقد يُكتب بجوار كل كلمةٍ وسياقها رقمٌ أو رمزٌ يقود المستفيد إلى المصدر الذي وردت به الكلمة. وبذلك نجد أن سطر تكشيف الكلمات المفتاحية في السياق يتكون من:

١. الكلمة المفتاحية: وهي الرأس أو المدخل في تسلسلها الهجائي.
٢. السياق: وهو يتمثل في الكلمات المحيطة بالكلمة المفتاحية. ويفيد هذا في التحقق من البيئة اللغوية التي وردت بها الكلمات.
٣. الكود: وهو الإحالة أو الإشارة المرجعية التي تربط المداخل بالبيانات الببليوجرافية الكاملة للنصوص المكشفة، أو يحدد أماكن وجودها (عبدالهادي، ١٩٨٢، ص ٧٥).

ونظرًا إلى أن الاكتفاء بتكشيف العناوين لا يكفل تعبيرًا دقيقًا لمحتوى المدونات اللغوية، فقد بات من الضروري المزج بينه وبين تكشيف النصوص. وقد أسهمت التطورات التقنية الحديثة إلى تحقيق تقاربٍ بين نمطي التكشيف هذين، حتى أصبحت أدوات معالجة المدونات اللغوية تُزاوج بين هذين النمطين؛ فما يتم في الأول من معالجةٍ يتم في الثاني، وما يحدث في الثاني يحدث أيضًا في الأول (عبد الهادي، ١٩٨٢، ص ٢٩).

ولذا نجد أن بعض الدراسات التي تتناول المدونات اللغوية أضحت تُطلِق على هذه المعالجة الفنية للكلمات تكشيفَ النصوص فقط، أو تكشيفَ الكلمات المفتاحية في السياق فقط، وبعض الدراسات الأخرى رادفت بين النمطين.

المشكلات الفنية في تكشيف المدونات اللغوية

على الرغم مما يتميز به هذان النمطان من التكشيف إلا أن ثمة نقطتي ضعف تكتنفهما، وهما:

١. تشتت مداخل الموضوع الواحد بسبب افتقاد الكلمات للتقنين أو التوحيد. حيث تتشتت الوثائق المتصلة بموضوعٍ معينٍ تحت الأشكال المختلفة للتعبير عن هذا الموضوع.
٢. بتر بعض الكلمات بسبب تحديد طول السطر بعددٍ معينٍ من الأحرف؛ الأمر الذي يؤدي عادةً إلى ضياع جزءٍ من السياق.

المشكلات اللغوية في تكشيف المدونات اللغوية

يُضاف إلى المشكلات الفنية السابقة بعضُ المشكلات الأخرى المرتبطة باللغة العربية، منها:

١. الترادف: الذي يؤدي إلى تشتت الموضوعات تحت الكلمات المترادفة الدالة على مفهومٍ أو معنىً واحدٍ (عبد الهادي & زايد، ٢٠٠٠، ص ص ٣٧-٣٩).
٢. الجناس: الذي يتطلب عرض السياق كاملًا كي يمكن التحقق من المعنى الصحيح المراد للكلمة.
٣. اختلاف الإملاء: الذي يؤدي أيضًا إلى تشتت الموضوعات تحت الرسم المختلف للكلمات، وبصفةٍ خاصةٍ الكلمات الأجنبية والمُعرَّبَة، مثل: جوجل/ غوغل/ قوقل، جرام/ غرام (عبد الهادي & زايد، ٢٠٠٠، ص ص ٣٧-٣٩، ٨٤-٨٥).

قوائم تردد الكلمات

من الوظائف الأخرى التي تقوم عليها برمجيات المدونات اللغوية ترتيبُ الكلمات وفقًا لتكرار ترددها بالنصوص. ويُعد مدى تكرار تردد الكلمات في المدونات اللغوية مؤشرًا على أهمية هذه الكلمات لأغراض التعبير عن المحتوى. حيث يتم إحصاء عدد مرات تردد الكلمات في المدونة اللغوية وترتب تنازليًا. وتُعد الكلمات التي ترد على

القمة (أكثر الكلمات ترددًا) هي الكلمات المحورية[1] في مجموعةٍ ما من النصوص، أو في لغةٍ ما، أو في تخصصٍ ما... ومن ثم يمكن التحقق من أشكال هذه الكلمات، بغرض اعتماد إحداها، أو ربما إغفالها.

أما إذا ما تم استبعاد الكلمات الوظيفية من قمة الترتيب، فإن ما يتبقى من كلماتٍ يمكن اختياره ليكون مصطلحاتٍ كشفيةً. وكذلك يمكن إحصاء تكرار المقاطع، والجمل، والعبارات، والفقرات.

وقد يتم ترتيب الكلمات ترتيبًا ألفبائيًا، وقد يتم وفقًا لترتيب ظهور الكلمات داخل النص كونها دليلًا سريعًا لتوزيعات الكلمات داخل النصوص (Lijffijt, 2011, p. 342).

توليد الكلمات المفتاحية

تعتمد هذه العملية الفنية بشكلٍ أساسٍ على سابقتها؛ وذلك عن طريق المقارنة بين قوائم تردد الكلمات. فالكلمة يتم احتسابها مفتاحيةً إذا كان تكرار ترددها عاليًا (أو ربما منخفضًا) على نحوٍ غير معتادٍ؛ وذلك في إطار مقارنتها بمدونةٍ لغويةٍ مرجعيةٍ reference corpus.

حيث تُستخدم عادةً قائمتان (على الأقل) لتكرار تردد الكلمات من أجل توليد الكلمات المفتاحية لمجموعةٍ من النصوص؛ إحداهما تُمثل النصوص المستهدفة التي يتم دراستها، والأخرى تُمثل النصوص المرجعية، والتي تكون أكبر من الأولى، وهي في الوقت نفسه تُمثل خلفيةً للبيانات التي يتم احتساب الكلمات المفتاحية من خلالها (Culpeper, 2009, p. 33).

حيث تتم مضاهاة الكلمات الواردة في القائمتين، وتُطَبَق بعض الطرق الإحصائية، والتي منها: اختبار دوال الاحتمالات Likelihood Ration Functions، أو اختبار مربع كاي Chi-Square (Scott, 2009). وتكون النتيجة قائمةً بالكلمات المفتاحية، سواءٌ ذات التردد العالي غير المعتاد (كلماتٌ مفتاحيةٌ إيجابيةٌ) أو تلك ذات التردد المنخفض غير المعتاد (كلماتٌ مفتاحيةٌ سلبيةٌ).

إلا أنه ينبغي مراعاة عاملين رئيسين قبل اعتماد مدونة النصوص المرجعية لعملية المقارنة؛ وهما:

١- ليس شرطًا أن تكون كلماتٍ دالةً؛ لأن أكثر الكلمات ترددًا تكون عادةً هي الكلمات الوظيفية.

١. حجم هذه المدونة: فعلى الرغم من أن البعض قد قلل من أهمية الدور الذي يؤديه حجم المدونة المرجعية لتردد الكلمات (McEnery, 2009)، إلَّا أن البعض الآخر أكد على ضرورة أن يبلغ حجم المدونة المرجعية لتكرار تردد الكلمات خمسةَ أضعاف المدونة المدروسة، وإلَّا فإن النتائج ستكون غير دقيقةٍ (Berber-Sardinha, 2000, pp. 7-8). فيما أوصت بعض الأدبيات الخاصة بموضوع استخراج الكلمات المفتاحية بأن يكون حجم المدونة المرجعية أكبر من حجم المدونة المدروسة، أو على أقل تقدير أن تكون المدونتان متساويتين في الحجم (الثبيتي، ٢٠١٦، ص ١٠٥).

٢. نوع هذه المدونة: ليس هناك اختلافٌ ذو دلالةٍ إحصائيةٍ بين النصوص المكتوبة والنصوص المنطوقة حال اعتمادها أنواعًا للمدونات المرجعية لتكرار تردد الكلمات. إلا أن تصنيف هذه المدونات في فئاتٍ (أخبار، أو ثقافة، أو أدب... إلخ) هو العامل المؤثر إيجابًا في استخراج وتوليد الكلمات المفتاحية. فإذا ما أُريد توليد كلماتٍ مفتاحيةٍ لقائمة تردد كلماتٍ لنصوصٍ أكاديميةٍ، فإنه يُفضَّل أن تكون المدونة المرجعية المستخدَمة لنصوصٍ إخباريةٍ مثلًا، أو ربما نصوصٍ أدبيةٍ... أو غير ذلك.

يُذكر أنه يمكن الاستفادة من أسلوب المقارنة هذا، والطرق الإحصائية المستخدمة به، في توليد قائمةٍ بالكلمات المفتاحية المستخدَمة في فترةٍ زمنيةٍ معينةٍ؛ وذلك من خلال المقارنة بين قائمة تردد كلماتٍ لإحدى الفترات، في مقابل مدونةٍ مرجعيةٍ لتردد كلماتٍ من فترةٍ زمنيةٍ أخرى. وقد تُستخدم أيضًا هذه الطريقة للحصول على قائمةٍ بالكلمات المفتاحية المستخدَمة في نوعٍ لغويٍ variety معينٍ؛ كما هو الحال عند المقارنة بين الكلمات الأكثر ترددًا في الفصحى مقابل العامية. أو في فرعٍ معينٍ من فروع المعرفة البشرية مقابل فرعٍ آخر... إلخ (Goh, 2011, pp. 251-54).

تحليل التجمعات العنقودية Cluster analysis

إلى جانب أنه يمكن للبرمجيات المعالِجة للمدونات اللغوية أن تقوم بفرز وترتيب الكلمات الأكثر ترددًا، فإنه يمكنها أيضًا أن تقوم بالتحليل والترتيب وفقًا لمجموعات الكلمات combinations of words أو المقاطع chunks؛ وذلك بأن يتم تحديد عدد الكلمات المسترجَعة: 2-word combinations أو 3- أو 4-... فإذا تم البحث في المدونة

اللغوية عن المقاطع المكوَنة من ثلاث كلماتٍ، فإنه سيتم النظر أولًا إلى الكلمة ١ و ٢ و ٣ التي يبدأ بها النص، ثم يتم النظر إلى الكلمة ٢ و ٣ و ٤، ثم إلى الكلمة ٣ و ٤ و ٥، وهكذا. وفي النهاية تُعرَض صفحة نتائج عبارةٌ عن قائمةٍ من تجمعاتٍ عنقوديةٍ/ مقاطع لثلاث كلماتٍ three-word clusters/chunks تتكرر عددًا من المرات المقطعية cut-off points، وليكن مثلًا: أكثر من مرتين، أو أكثر من ثلاث مراتٍ، أو غير ذلك، وفقًا لرغبة المستفيد المسْبقة. مع الوضع في الحسبان أن آلية التحليل تحتسب أي عددٍ من الحروف (التمثيلات) بعدها (أو قبلها) مسافةٌ على أنها كلمةٌ واحدةٌ؛ بمعنى أنه إذا كان هناك اسمٌ أو فعلٌ يرتبط به ضميرٌ، فإنه سيتم التعامل معه كونها كلمةً واحدةً، وليس كلمتين.

ومن هنا نجد أن العبارات المسترجَعة من تحليل التجمعات العنقودية تفتقد غالبًا إلى الاكتمال النحوي؛ مثل: وبطريقةٍ أخرى... ويمكن أن... لابد أن... مجموعةٌ من... إلا أنها قد تؤدي وظيفةً تداوليةً. pragmatics (O'Keeffe; McCarthy & Carter, 2007, pp 70-71). فمثلًا المقطع «بإذن الله...» قد يعبر في سياقاتٍ معينةٍ عن مواقف إيجابيةٍ (التمني)، إلا أنه في سياقاتٍ أخرى قد يعبر عن مواقف سلبيةٍ (النفي).

وعلى الرغم من ذلك، فإن تحليل التجمعات العنقودية يسهم في التحقق من الوظائف أو المقيدات النحوية syntactic restrictions للكلمة أو للكلمات. إذ يمكن التأكد من حروف الجر التي تدخل عادةً على بعض الكلمات وتُغيِّر أو تُضيف إلى معناها، والمكان النحوي المتوقع لها (في بداية الجملة، أم في منتصفها، أم في أخرها) (O'Keeffe; McCarthy & Carter, 2007, pp. 14-16).

ويُسهم تحليل التجمعات العنقودية أيضًا في التحقق من المقيدات الدلالية للكلمات؛ حيث يمكن تحديد ما إذا كانت الكلمة تُستخدم مع العاقل فقط، أم مع غير العاقل فقط، أم مع كليهما.

كما أن العبارات المسترَجعة قد تؤدي وظيفةً تُعرف بـــ «العَروض الدلالي semantic prosody» وقد استخدم هذا المصطلح لوو Louw (1993)؛ ويعني أن ترد الكلمة مع كلمةٍ أو كلماتٍ أخرى يُكسبها دلالةً معينةً في سياقاتٍ محددةٍ بالإيجاب أو بالسلب (O'Keeffe; McCarthy & Carter, 2007, pp. 14-16).

وفضلًا عن ذلك، فإن تحليل التجمعات العنقودية ينصب عادةً على الكشف عن الكلمات التي تتوارد لتؤلِف عباراتٍ بلاغيةً مكوَنةً عادةً من كلمتين، وأحيانًا من

ثلاث أو أكثر تتصاحب مع بعضها عادةً وتتلازم في اللغة، وتسمى حينئذٍ «متلازماتٌ لغويةٌ» أو «متصاحباتٌ لغويةٌ»[1] collocates، مثل: رأب الصدع (كلمتان)، كالنار في الهشيم (أربع كلمات)، فتح الباب على مصراعيه (خمس كلمات)... وهي ليست عباراتٍ مفتوحةً مطاطةً، بل هناك عُرفٌ سائدٌ في استعمالها يجعل معظمها عباراتٍ شبه ثابتةٍ يستخدمها المتخصصون وغيرهم بديهيًا.

وفي مجال المكتبات والمعلومات، على سبيل المثال، يمكن لتحليل التجمعات العنقودية أن يسهم في الكشف عن متلازماتٍ لغويةٍ مثل: أوعية المعلومات، وتنمية المقتنيات، ونظم استرجاع المعلومات... وغيرها؛ فحينما ترد كلمة «أوعية» في مدونةٍ لغويةٍ لمجال المكتبات والمعلومات، فإننا نعلم أنها متصاحبةٌ مع كلمة «المعلومات»، أما إذا وردت في مدونةٍ لغويةٍ للعلوم الطبية، فإنها بالضرورة ستتصاحب مع كلمة «دموية».

وبذلك فإن تحليل التجمعات العنقودية للمدونات اللغوية يفيد كثيرًا في دراسة وتحليل الكلمات وصك المصطلحات (صناعة المعاجم والمكانز)، فضلًا عن أنه يكشف عن التعبيرات الاصطلاحية idiomatic expressions، وأسلوب الاتصال، أو الأسلوبية stylistics، والأسلوب العِباري للكلمات phraseology الذي يفيد في التحقق من الخصائص اللغوية المميزة لتخصصٍ ما، أو لمؤلفٍ ما (غزالة، ٢٠٠٧، ص ص ٥-١٢). وهو بذلك أسهم بشكلٍ أو بآخر في أن جعل الكثير من المتخصصين ينظرون إلى كون الكلمات أكبر من مجرد «قائمةٌ غير مرتبةٍ من وحداتٍ معجميةٍ مفردةٍ» كما ذهب تشومسكي (Chomsky, 1965, p. 84).

١- هي كلماتٌ تتصاحب أو تتلازم أو تَقترن دائمًا بكلماتٍ أخرى، سواءٌ لسببٍ، أو دون سببٍ ظاهرٍ أو منطقيٍ.

الخلاصة

تناول هذا الفصل من الكتاب المدونات اللغوية من حيث كيفية معالجة البيانات اعتمادًا عليها. وقد شملت عمليات المعالجة تلك المعالجات التي تتم على مستوى الإدخال: ترميز المدونات اللغوية، وتحشية المدونات اللغوية. وتعد الأخيرة واحدةٌ من أبرز العمليات الفنية التي تتم على المدونات اللغوية. حيث تمت مناقشة هذه العملية الفنية في ضوء تعريفها، وأهميتها، والنقد الموجه لها، وطرائقها، وأنواعها المختلفة التي تشمل: وسم أقسام الكلم، وتجريد الكلمات، والتحليل الإعرابي، والتحشية الدلالية، وتحشية المصاحبة المرجعية، والتحشية البراجماتية أو التداولية، والتحشية الأسلوبية، ووسم الأخطاء، والتحشية الموجهة نحو المشكلات، والتحشية المتضمَنة والتحشية القائمة بذاتها.

كما شملت عمليات المعالجة التي تم التركيز عليها في هذا الفصل تلك التي تتم على مستوى الإخراج: تكشيف الكلمات، وتكشيف النصوص، وتكشيف الكلمات المفتاحية في السياق، والمشكلات الفنية في تكشيف المدونات اللغوية، والمشكلات اللغوية في تكشيف المدونات اللغوية، وقوائم تردد الكلمات، وتوليد الكلمات المفتاحية، وتحليل التجمعات العنقودية.

الفصل الرابع

المدونات اللغوية: نماذج وبرمجياتٌ

التمهيد
المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية
المدونة اللغوية التاريخية للجامعة الأردنية
المدونة اللغوية العربية الدولية لمكتبة الإسكندرية
مدونة عربي كوربَص
المدونة اللغوية لمتعلمي اللغة العربية
المدونة العربية القرآنية
مدونة قرآني
استقصاء المدونات اللغوية العربية
سكتش إنجين
المدونة اللغوية للإنجليزية الأمريكية المعاصرة
المدونة اللغوية للأخبار على الشبكة العنكبوتية
المدونة اللغوية لكتب جوجل
برمجيات معالجة وتحليل المدونات اللغوية العربية
برنامج أدوات وورد سميث
برنامج أدوات معالجة المدونات اللغوية العربية

التمهيد

تناولنا فيما سبق من هذا الكتاب المدونات اللغوية؛ تلك الأداة التي تؤدي دورًا بالغ الأهمية لكثير من المجالات العلمية. إذ تم تناول هذه الأداة في ضوء ماهيتها وأهميتها، ومميزاتها، ومواصفاتها، وما يمكن أن تقدمه من خدماتٍ بحثيةٍ. كما تم تناول العمليات الفنية التي تتم على المدونات اللغوية وأساليب معالجة نصوصها.

وبقي لنا هنا أن نؤكد أن المدونات اللغوية تُتاح عادةً من خلال واجهة تعامل على الشبكة العنكبوتية تُمكن الباحثين والمستفيدين من إجراء عددٍ من التحليلات الكمية للكلمات والمصطلحات الكشفية. وتغطي هذه المدونات اللغوية عادةً أيضًا موضوعًا محددًا، أو نوعًا أدبيًا بعينه، أو فئةً من مصادر المعلومات، أو غيرها خلال فترةٍ زمنيةٍ معينةٍ.

وفيما يلي مجموعةٌ من المدونات اللغوية القائمة على الشبكة العنكبوتية، مع تسليط الضوء على العربية منها التي يتضح فيها جليًا نظام تكشيف الكلمات المفتاحية في السياق، وعرض النتائج المسترجعة وفقًا لسياقات الكلمات المفتاحية التي تعبر عن موضوعاتٍ متعددةٍ في مختلف فروع المعرفة البشرية. حيث يتم تناول هذه المدونات اللغوية في ضوء حجمها، والمجالات والموضوعات التي تغطيها، وسماتها وخصائصها، وإمكانات البحث التي تقدمها.

وقد وضع المؤلف لنفسه، عند الحديث عن هذه المدونات اللغوية، نموذجًا معياريًا استلهمه من البيانات المتاحة حول هذه المدونات. حيث وجد أن هذه البيانات تغطي مجموعةً من حقول الوصف التي إلتزم بها المؤلف، قدر الإمكان، عند وصف المدونات، على النحو التالي:

١. الاسم.
٢. المحدد الموحد لمكان المصدر URL.
٣. التعريف والهدف.
٤. بيانات الإنشاء، وتشمل: المنشيء، وسنة الإنشاء.
٥. الحجم.
٦. اللغة والأنواع اللغوية varieties.
٧. العينة.
٨. نوع النصوص؛ مكتوبةٌ، أم منطوقةٌ.
٩. الأنواع الأدبية genres والموضوعات.
١٠. الفترة الزمنية المغطاة.
١١. المصادر اللغوية/ مصادر المعلومات والمصادر المكانية.
١٢. التحشية والترميز.
١٣. إمكانات التحليل.
١٤. لقطة مصورة من شاشة أو شاشات عرض المدونة اللغوية.

يعقب ذلك استعراضٌ لبعض برمجيات معالجة المدونات اللغوية العربية، مع إبراز إمكاناتها، والنظم الفرعية التي تتضمنها.

المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية

King Abdulaziz City for Science and Technology (KACST) Arabic Corpus

هي مدونةٌ لغويةٌ عربيةٌ متاحةٌ بالمجان من خلال هذا الرابط: corpus.kacst.edu.sa /index.jsp . وتفيد هذه المدونة اللغوية في أغراضٍ بحثيةٍ مختلفةٍ؛ تبدأ من الدراسات اللغوية بمستوياتها المتنوعة، وتمتد لتشمل تطوير تطبيقات معالجة اللغة الطبيعية (Al-Thubaity, 2015).

وقد بادرت مدينة الملك عبد العزيز للعلوم والتقنية إلى إنشاء هذه المدونة في عام ٢٠١٢. ويتجاوز حجم المدونة المليار كلمةٍ (١٨٢٥١٥٦٣٣ كلمةٍ)؛ وبذلك فهي تُعد واحدةً من أكبر المدونات العربية. وتمثل المدونةُ النصوصَ المكتوبة للفصحى (الفصحى التراثية، والعربية القياسية الحديثة) فقط، دون اللهجات العربية المختلفة. وذلك عبر ثلاثة أبعادٍ رئيسةٍ، هي: الفترة الزمنية، والمنطقة الجغرافية، والنوع الأدبي. وتمتد الفترة الزمنية المغطاة من عصر ما قبل الإسلام، وحتى وقت إنشاء هذه المدونة اللغوية.

وتم جمع نصوص المدونة من مصادر معلوماتٍ متنوعةٍ، اعتمادًا على ما هو متوافر على الشبكة العنكبوتية من محتوىً عربيٍ. وتشمل ١٠ مصادر معلوماتٍ، هي: المخطوطات المحققة، والكتب، والصحف، والمجلات، والمناهج الدراسية، والرسائل العلمية، والمواقع الإلكترونية، والدوريات المحكمة، والمطبوعات الرسمية، ووكالات الأنباء. وتم تصنيف هذه النصوص على ٨٠ نطاقًا، و ٤٨١ موضوعًا.

وتم الحصول على مصادر المعلومات هذه من مصادر لغويةٍ مختلفةٍ. فعلى سبيل المثال، تم الحصول على المخطوطات القديمة من موقع المكتبة الشاملة. وتم تجميع نصوص الكتب من موقع المكتبة الشاملة، وموقع صيد الفوائد، وموقع اتحاد الكتاب العرب. أما نصوص الدوريات العلمية المحكمة فقد تم تجميعها من مواقعها الإلكترونية لعددٍ من الجامعات العربية، كجامعة أم القرى، وجامعة الملك فيصل، إضافة إلى موقع اتحاد الكتاب العرب. فيما تم تجميع نصوص الرسائل العلمية من مواقع عددٍ من الجامعات العربية، وموقع المكتبة الشاملة.

وجُمِّعت نصوص المطبوعات الرسمية من مواقع قانونية متخصصة، مثل مواقع وكالات الأمم المتحدة، والمواقع الحكومية. بينما جُمِّعت نصوص المناهج الدراسية من مواقع الجامعات، ومواقع وزارات التربية والتعليم، والمواقع التعليمية. وجُمِّعت نصوص الصحف، والمجلات، ووكالات الأنباء من موقع صحيفة الوطن السعودية، وموقع صحيفة روز اليوسف المصرية، ووكالة الأنباء السعودية، وغيرها.

وتوزعت الأنواع الأدبية والموضوعات التي تشملها المدونة على ما يتناسب مع الفترة الزمنية التي تغطيها. فعلى سبيل المثال، نجد أن الأخبار الرياضية مناسبةُ للصحف، بوصفها المصدر اللغوي، في الفترة الزمنية الحديثة فقط، وهكذا. ويظهر الشكل رقم

(١) الصفحة الرئيسة للمدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية.



الشكل رقم (١) الصفحة الرئيسة للمدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية

وراعت هذه المدونة اللغوية حقوقَ الملكية الفكرية، لاسيما مع المصادر الحديثة التي استقت منها النصوص، وخاصة الصحف، ووكالات الأنباء، والمجلات، والمواقع الإلكترونية. حيث وضعت المدونة قيودًا للإتاحة يتم بمقتضاها مراعاة حقوق الملكية الفكرية. وتمثلت هذه القيود في: عدم توزيع النصوص التي تم جمعها، وعدم السماح بتحميل هذه النصوص، وعدم السماح بعرض النصوص الكاملة لها، ولكن عرض السياقات التي وردت بها الكلمات فقط (١٥ كلمةً قبل وبعد الكلمات المحورية). وعلاوةً على ذلك، فقد وفر القائمون على هذه المدونة البيانات الببليوجرافية الخاصة بها. وذلك كله في إطار إعفائها من أي انتهاكٍ للحقوق، واتساقها، في الوقت نفسه، مع القانون السعودي للملكية الفكرية.

وتتميز المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية بتعزيز نصوصها بالميتاداتا وفق مجموعة من حقول الوصف، هي: العنوان، وتاريخ النشر، والفترة الزمنية، واسم المؤلف ونوعه (ذكر أم أنثى)، والمنطقة، ومصدر اللغة، والنطاق، والموضوع. الأمر الذي يتيح للباحثين والمستفيدين من المدونة إمكانية الاسترجاع والتحليل وفق مقيدات بحثٍ مختلفةٍ.

وتوفر المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية عددًا من إمكانات الاسترجاع، هي:

١. **البحث: توفر هذه الأداة نوعين من البحث:**

البحث في نصوص المدونة بكلمةٍ واحدةٍ أو عدةٍ كلماتٍ بطريقةٍ مشابهةٍ لمحركات البحث. حيث تظهر السياقات التي وردت فيها هذه الكلمات مع معلوماتٍ تحوي عنوان النص، والوعاء، والمجال، والموضوع، والفترة الزمنية، والبلد الذي طبع فيه النص. ويمكن للمستفيد تحديد البحث في جزءٍ، أو عدة أجزاءٍ محددةٍ من المدونة من خلال المحددات الموجودة في أعلى الصفحة. ما يمكنه تحديد أكثر من عنصرٍ من عناصر المحدد الواحد بالضغط على زر CTRL من لوحة المفاتيح واختيار العناصر المطلوبة. ويمكن تطبيق هذه الطريقة على جميع محددات البحث في الوقت نفسه. ويجدر بالذكر أن هذه الخاصية متوفرةٌ أيضًا في بعض الأدوات (الكشاف السياقي، والتصاحب اللفظي).

البحث عن عناوين النصوص بواسطة كلمة أو مجموعة كلمات. وهي مشابهة للبحث في النصوص لكن تبحث في عناوين النصوص فقط حيث تظهر النصوص التي وردت في عناوينها هذه الكلمات مع معلوماتٍ تحوي النص، والوعاء، والمجال، والموضوع، والفترة الزمنية، والبلد الذي نشر فيه النص. ويوضح الشكل رقم (٢) صفحة البحث في المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية.



الشكل رقم (٢) صفحة البحث في المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية

٢. **البحث المخصص**: تمكن هذه الأداة المستفيدَ من البحث عن الكلمات بواسطة: الجذع؛ بحيث يظهر الجذع بصوره المختلفة بعد إضافة السوابق واللواحق الممكنة (مثل أل التعريف، والواو، والباء، والضمائر المتصلة) وذلك بحسب نوع الجذع (اسم، أو فعل) مع معلومات عن تكرار هذه الصور في المدونة. ويمكن من خلال النقر على الكلمة الانتقال إلى الكشاف السياقي للكلمة.

البحث بواسطة رموز البدل (٪ و ـ) حيث تعني (٪) أي عددٍ من الأحرف، وتعني (ـ) حرفًا واحدًا فقط. ويمكن أن يوضع أي رمزٍ من هذين الرمزين في أي مكان من سلسلة الأحرف التي يراد البحث عنها. فعند إدخال السلسلة (٪م٪) فيعني هذا البحثَ عن أي كلمةٍ تحتوي على الحرف م. وتعني السلسلة (م٪) أي كلمةٍ أولها حرف الميم. وتعني السلسلة (ــ م ــ) البحثَ عن أي كلمةٍ مكونةٍ من ثلاثة حروف أوسطها حرف الميم، وهكذا. ويبين الشكل رقم (٣) صفحة البحث المخصص في المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية.



الشكل رقم (٣) صفحة البحث المخصص في المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية

٣. **توزيع التكرار**:

يمكن من خلال هذه الأداة معرفة التوزيع الإحصائي لكلمةٍ واحدةٍ، أو كلمتين متتابعتين، أو ثلاث حسب الفترات الزمنية، أو حسب أوعية المدونة؛ كما يختار

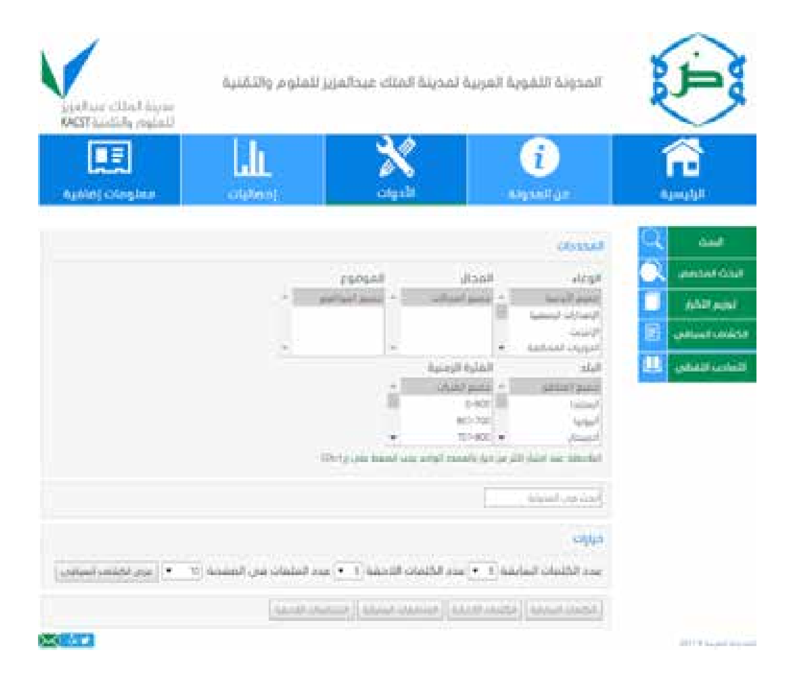
المستفيد. حيث تظهر الفترة الزمنية، أو الوعاء، ويقابلها تكرار الكلمة، وتكرارها النسبي، وعدد النصوص التي وردت فيها. ويظهر الشكل رقم (٤) صفحة توزيع التكرار في المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية.



الشكل رقم (٤) صفحة توزيع التكرار في المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية

٤ . الكشاف السياقي:

تستخدم هذه الأداة لعرض السياقات التي تظهر فيها الكلمة، أو أي مجموعةٍ من الكلمات المتتابعة ضمن مدى محددٍ من الكلمات السابقة واللاحقة. ويتراوح عدد الكلمات السابقة أو اللاحقة من صفر وحتى ١٥ كلمةً. وتفيد هذه الأداة بشكلٍ رئيسٍ في الكشف عن دلالة الكلمة المركزية للسياق، وطبيعة السياقات التي تظهر فيها هذه الكلمة. ويجدر بالذكر أنه يمكن للمستفيد حفظ هذه النتائج في حاسبه الآلي من خلال الضغط على زر الحفظ الموجود في أعلى الصفحة أو أسفلها. كما يمكنه التحكم في عدد النتائج التي تظهر في الصفحة. ويوضح الشكل رقم (٥) صفحة الكشاف السياقي في المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية.



الشكل رقم (٥) صفحة الكشاف السياقي في المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية

٥. التصاحب اللفظي:

تستخدم هذه الأداة لقياس مدى الارتباط بين الكلمة المبحوث عنها، والكلمات التي تظهر معها في السياق حسب طول السياق الذي يحدده المستفيد من خلال الكلمات السابقة والكلمات اللاحقة. ويمكن للمستخدم اختيار طريقةً إحصائيةً أو أكثر من ثمانِ طرقٍ ممكنةٍ في النظام لقياس قوة الترابط. ويظهر النظام المعلومات التالية للمستفيد: الكلمة وتكرارها في المدونة، وتكرارها في السياق، وقيمة معامل الارتباط الإحصائي. ويمكن للمستفيد تصفيةَ النتائج والاقتصار على كلماتٍ معينةٍ بحسب تكرار هذه الكلمات في كامل المدونة و/ أو السياق. كما يمكن أيضًا تصفية النتائج والاقتصار على الكلمات التي تحقق قيمةً معينةً على الأقل حسب نتائج المقاييس الإحصائية. ويمكنه كذلك ترتيب النتائج تصاعديًا أو تنازليًا بالنقر على رأس العمود الذي يريد ترتيب النتائج بناءً عليه. كما يمكنه حفظ هذه النتائج في حاسبه الآلي من خلال الضغط على

زر الحفظ الموجود في أعلى الصفحة أو أسفلها. ويبين الشكل رقم (٦) صفحة التصاحب اللفظي في المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية.



الشكل رقم (٦) صفحة التصاحب اللفظي في المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية

المدونة اللغوية التاريخية للجامعة الأردنية Historical Arabic Corpus

هذه المدونة متاحةٌ عبر الإنترنت من خلال هذا الرابط: nlp.ju.edu.jo . وتهدف هذه المدونة إلى خدمة علماء اللغة ومتعلمي العربية بحيث يمكنهم استكشاف وفهم الاستعمال اللغوي وتطوره، والتحقق من التغير الدلالي عبر العصور والمراحل الزمنية المختلفة للأدب العربي (Hammo, Yagi, Ismail & AbuShariah, 2016). ويبلغ حجم هذه المدونة ٤٥ مليون هيكل كلمةٍ من مختلف العصور التاريخية للأدب العربي. وذلك من خلال ما تحويه المدونة من نصوصٍ عربيةٍ قديمةٍ تمتد لأكثر من ستة عشر قرنًا من الاستعمال اللغوي؛ منذ العصر الجاهلي، وعصر صدر الإسلام، والعصر العباسي

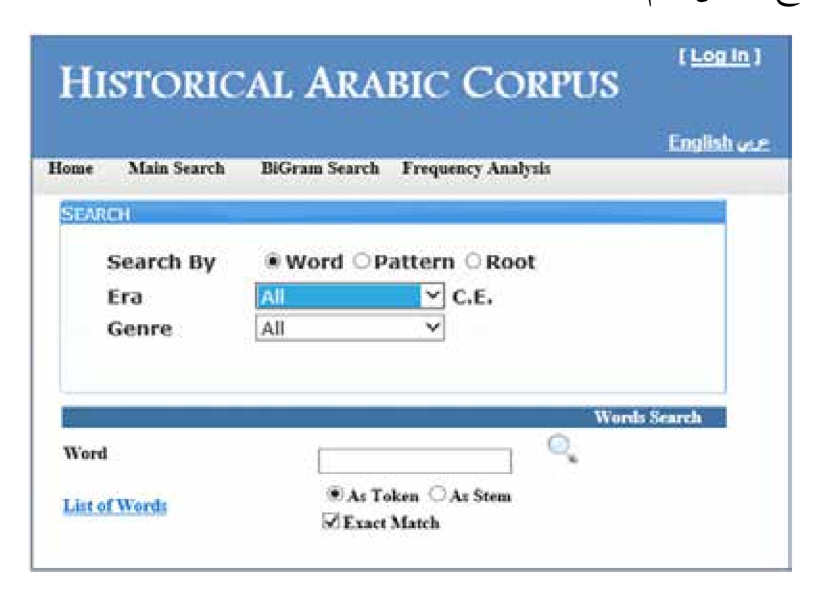
الأول، والعصر العباسي الثاني، وعصور الولايات والممالك المتتالية، وحتى العصر الحديث. وكانت الشبكة العنكبوتية هي المصدر الأساس للحصول على هذه النصوص. وتشمل المدونة أنواعًا أدبيةً عدةً مصنفةً على النحو التالي: نثر، وشعر، وتاريخ، وفلسفة، ودين، وعلوم، ومعتقدات، ومعاجم لغوية.

ويُقر القائمون على هذه المدونة بأنها تفتقد كثيرًا إلى التمثيل الجيد للبيانات، والتوازن بين المتون؛ بسبب الصعوبة الكبيرة في الحصول على النصوص العربية المرقمنة.

كما تمت معالجة كل كلمةٍ من كلمات المدونة (فيما عدا الكلمات المستثناة) بحيث يتم تحديد الكلمة نفسها، وموقعها داخل النص، والجذر، والوزن، وقسم الكلم، والجذع.

وتم ترميز بيانات المدونة اللغوية التاريخية للجامعة الأردنية باستخدام لغة الترميز القابلة للتمديد XML متضمنةً مجموعة من حقول الميتاداتا، هي: عنوان الوثيقة، والمؤلف، وتاريخ النشر، والعصر أو الفترة الزمنية، والفئة، والنوع الأدبي، والمنشىء، تاريخ الإنشاء، وتاريخ آخر تعديل.

وتتيح هذه المدونة اللغوية إمكانات استرجاع البيانات وفق قائمة تردد الكلمات، وتركيب الجمل والعبارات، وعرض المتلازمات اللغوية لكل فترةٍ زمنيةٍ أو نوعٍ أدبيٍ. ويوضح الشكل رقم (٧) الصفحة الرئيسة للمدونة اللغوية التاريخية للجامعة الأردنية.



الشكل رقم (٧) الصفحة الرئيسة للمدونة اللغوية التاريخية للجامعة الأردنية

المدونة اللغوية العربية الدولية لمكتبة الإسكندرية
International Corpus of Arabic (ICA)

تُتاح هذه المدونة اللغوية على الشبكة العنكبوتية تحت اسم «المدونة اللغوية العربية العالمية» من خلال هذا الرابط: www.bibalex.org /ica /ar /about.aspx . وهي تمثل أحد المشروعات الثقافية التابعة لمكتبة الإسكندرية الهادفة لبناء مدونةٍ لغويةٍ للعربية المعاصرة تحوي ١٠٠ مليون هيكل كلمةٍ[1] مُحَلَّلَةً صرفيًا ونحويًا ودلاليًا، وممثلةً لقطاعٍ إقليميٍ كبيرٍ من الدول الناطقة باللغة العربية المعاصرة، وعاكسةً بشكلٍ حقيقيٍ وواقعيٍ لأنماط استخدام اللغة العربية المعاصرة في أنحاء الوطن العربي.

روعي في بناء هذه المدونة اللغوية التمثيلُ الجيد للنصوص في العربية المعاصرة، والتنوع في فئات النصوص ومحتواها، والتوازن بين كل فئةٍ من النصوص، وحجم الكلمات المُجَمَّعة في كل فئةٍ من فئات التجميع. وشملت المدونة اللغوية عددًا من المصادر والفئات المختلفة للنصوص؛ وذلك بهدف تحقيق شروط التمثيل الجيد، ومدى انتشار المصدر أو الفئة، والتوازن بين كل مصدرٍ وكل فئةٍ، وحجم الكلمات في كل مصدرٍ وفئةٍ.

ويعتمد تصميم هذه المدونة اللغوية على البدء بحصر المصادر المختلفة، وداخل كل مصدرٍ تم إدراج الفئات المميزة له. ويتم حفظ النصوص داخل هذه المدونة اللغوية بطريقةٍ هرميةٍ من خلال فهرسة النصوص وفقًا لثلاثةٍ من حقول الوصف، وهي: المصدر، والفئة، وتاريخ النشر.

ومن الأمور التي تم وضعها في الحسبان عند تجميع هذه المدونة اللغوية عددُ الفئات المتضمَنة داخل المدونة اللغوية، وعددُ النصوص داخل كل فئةٍ من هذه الفئات، بالإضافة إلى متوسط عدد الكلمات داخل كل نصٍ تبعًا لطبيعة كل مصدرٍ من مصادر التجميع.

وتوجد أربعة مصادر رئيسةٍ تقوم عليها هذه المدونة اللغوية، هي: الصحافة (٢٩٪)، والمقالات الإلكترونية (٢٠٪)، والكتب (٤٣٪)، والدراسات الأكاديمية (٨٪). وينقسم المصدر الخاص بالصحافة إلى ثلاثة مصادر فرعيةٍ، هي: الجرائد، والمجلات، والصحافة الإلكترونية. ويوجد إحدى عشرة فئةً على مستوى المدونة اللغوية، هي: العلوم الاستراتيجية، والعلوم الاجتماعية، والرياضة، والدين، والأدب، والعلوم الإنسانية، والعلوم الطبيعية، والعلوم التطبيقية، والفنون، والثقافة، والسير الذاتية، والنصوص المتنوعة.

١- سلسلةٌ من الحروف، أو التمثيلات، أو كلتيهما، يسبقها فراغٌ ويتبعها فراغٌ.

كما يوجد أربعٌ وعشرون فئةً فرعيةً، هي: السياسة، والقانون، والاقتصاد، والاجتماع، والدين الإسلامي، والدين المسيحي، والأديان أخرى، والدين المقارن، والقصص، والشعر، والنثر، والدراسات اللغوية والأدبية، والطب، والهندسة، والزراعة، والتكنولوجيا، وعلم الأحياء، وعلم الفيزياء، وعلم الفضاء، وعلم الجيولوجيا والبيئة، وعلم الكيمياء، وعلم النفس، وعلم الفلسفة، والتاريخ.

ويوجد أربع فئاتٍ فرعيةٍ من فئة القصص الفرعية، هي: الروايات، والقصص القصيرة، وقصص الأطفال، والمسرحيات.

هذا، ويتم تحليل المدونة اللغوية بطريقةٍ آليةٍ مبنيةٍ على بعض الطرق الإحصائية وبعض القواعد اللغوية بالاعتماد على أحد المحللات الصرفية الشهيرة؛ وهي مُجَرِد بَكوالتر Buckwalter stemmer. حيث يتيح التحليل عددًا من إمكانات البحث، كالسوابق واللواحق، وأقسام الكلمات، وساقها، وجذعها، وجذرها، ووزنها الصرفي، بالإضافة إلى نوع الكلمة من حيث النوع (ذكر/أنثى)، والعدد، والتعريف تبعًا للسياقات المختلفة للكلمات داخل النصوص. ويبين الشكل رقم (٨) الصفحة الرئيسة للمدونة اللغوية العربية الدولية (المدونة اللغوية العربية الدولية لمكتبة الإسكندرية، 2013؛ Alansary, Nagi & Adly, 2008؛ Alansary, S. & Nagi, M. 2014).



الشكل رقم (٨) الصفحة الرئيسة للمدونة اللغوية العربية الدولية

مدونة عربي كوربَص arabiCorpus

هي مدونةٌ لغويةٌ عربيةٌ متاحةٌ عبر الرابط التالي: arabicorpus.byu.edu . وكانت هذه المدونة بالأساس مشروعًا تم تصميمه وتطويره بواسطة ديلوورث باركينسون Dilworth Parkinson الباحث بجامعة بريجهام يانج الأمريكية Brigham Young University. وتتيح هذه المدونة اللغوية إمكانية استرجاع الكلمات والعبارات وفقًا لتكرار ترددها في عددٍ من الفئات. وتضم هذه الفئات خمسة أنواعٍ أدبيةٍ رئيسةٍ، هي:

١. الصحف Newspapers.
٢. الأدب الحديث Modern Literature.
٣. الأدب غير القصصي Nonfiction.
٤. العامية المصرية Egyptian Colloquial.
٥. الأدب قبل العصر الحديث Premodern.

وبعض هذه الفئات مقسمةٌ إلى فئاتٍ أخرى. فعلى سبيل المثال، تحتوي فئة الصحف على محتوىً لصحف المصري اليوم، والأهرام، وأعمدة الشروق، والغد، والحياة، والوطن، كلٌّ في مدونات لغويةٍ فرعيةٍ.

ويبلغ عدد كلمات مدونة عربي كوربَص ١٧٣٦٠٠٠٠٠ هيكل كلمةٍ موزعةٍ على فئاتها، يمكن البحث فيها بالحروف العربية أو بنقحرتها بالحروف اللاتينية. كما يمكن البحث وفقًا لنوع التحشية: اسمًا، أو فعلًا، أو صفةً، أو متواليةً string. ومن الممكن أيضًا البحث بإضافة التشكيل للكلمات أو بدونه، وبالهمزة أو بدونها. ويبين الشكل رقم (٩) الصفحة الرئيسة لمدونة عربي كوربَص.



الشكل رقم (٩) الصفحة الرئيسة لمدونة عربي كوربَص

المدونة اللغوية لمتعلمي اللغة العربية Arabic Learner Corpus

يقوم عليها عبد الله الفيفي، الباحث بجامعة الإمام محمد بن سعود الإسلامية، المملكة العربية السعودية. ويتمثل الهدف من هذه المدونة اللغوية (www.arabiclearnercorpus.com) في توفير نصوصٍ عربيةٍ مكتوبةٍ ومنطوقةٍ حررها دارسو اللغة العربية في المملكة العربية السعودية، على أن تكون هذه النصوص مفتوحة المصدر.

وتحتوي هذه المدونة اللغوية على مجموعةٍ من النصوص والتسجيلات في موضوعين مختلفين: الأول سردي (رحلةٌ خلال إحدى الإجازات)، والثاني للمناقشة (الاهتمامات الدراسية). وقد شارك في كتابة وتسجيل هذه المواد دارسو اللغة العربية في المملكة العربية السعودية خلال العامين ٢٠١٢ و ٢٠١٣.

وتشتمل هذه المدونة اللغوية على ٢٨٢٧٣٢ هيكل كلمةٍ مع التكرار (٣٨٦٥٧١ وحدةً صرفيةً)، و٢٩٦٢٧ كلمةً دون تكرار، عبارةٌ عن ١٥٨٥ مادةً (مكتوبةً أو مسجلةً) أنتجها ٩٤٢ طالبًا من ٦٧ جنسيةً و ٦٦ لغة أم مختلفة، تم تقسيمهم إلى مرحلتين: ما قبل الجامعة، والمرحلة الجامعية. ويبلغ متوسط طول النصوص في هذه المدونة اللغوية ١٧٨ كلمةً.

تم إنشاء هذه المدونة اللغوية لتزويد الباحثين بمجموعةٍ من البيانات مفتوحة المصدر؛ للاستفادة منها في مجالات البحث اللغوي، مثل: تعليم اللغة وتعلمها، وعلم اللغة التطبيقي، وصناعة المعاجم. كما يمكن استخدام هذه البيانات لأغراضٍ بحثيةٍ أخرى، مثل: تحليل الأخطاء اللغوية، وقياس التطور اللغوي لدى الطلاب، وتصميم المواد التعليمية، وتحليل اللغة المرحلية، وتأليف المعاجم الطلابية، وكذلك معاجم الأخطاء الشائعة.

وتقدم هذه المدونة اللغوية ثلاثة أنواع من البيانات:

١. بياناتٌ نصيةٌ في ملفات نوت باد Notepad بصيغة txt وبتشفير Unicode وأخرى بلغة الترميز القابلة للتمديد XML، وهذه تشمل جميع بيانات هذه المدونة اللغوية.
٢. صورٌ ممسوحةٌ ضوئيًا للمصدر الأصلي للنصوص المكتوبة يدويًا في ملفات بصيغة بي دي إف PDF، وهذه تشمل النصوص المكتوبة يدويًا فقط.

٣. تسجيلاتٌ صوتيةٌ (أكثر من ثلاث ساعاتٍ) في ملفاتٍ بصيغة MP3، وهذه تشمل التسجيلات الصوتية لأولئك الذين أعطوا الإذن بنشرها للاستخدام البحثي.

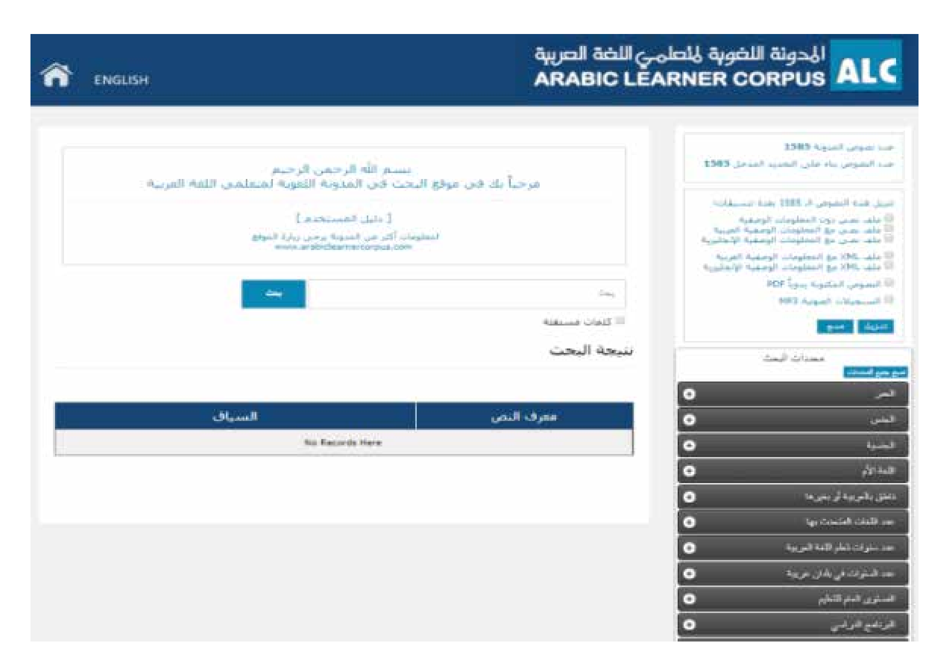
وجميع البيانات في هذه المدونة اللغوية متاحةٌ في نوعين من الملفات:

١. ملفاتٌ نصيةٌ:
   - بدون ميتاداتا metadata.
   - مع ميتاداتا باللغة العربية.
   - مع ميتاداتا باللغة الإنجليزية.
٢. ملفات بلغة الترميز القابلة للتمديد XML:
   - مع ميتاداتا باللغة العربية.
   - مع ميتاداتا باللغة الإنجليزية.

يمكن للباحثين من خلال الميتاداتا التحقق من خصائص النص اللغوي، وكذلك من قام بتحريره. الأمر الذي يضيف عمقًا أكثر لتحليل بيانات هذه المدونة اللغوية.

وتُتاح أصول الأوراق، المكتوبة يدويًا من قبل الطلاب، بعد أن تم إدخالها عن طريق الماسح الضوئي، وحفظها في ملفات بصيغة بي دي إف PDF. كما أن التسجيلات الصوتية (٣ ساعات، و٢٢ دقيقة، و ٥٩ ثانية) لأولئك الطلاب الذين أَذنوا بنشرها على الإنترنت، متاحةٌ للتنزيل في صيغة ملفات إم بي ثري MP3.

وتمت تسمية جميع ملفات هذه المدونة اللغوية بطريقةٍ تسمح بالتحقق من الخصائص الأساسية للنص وكذلك المؤلف؛ نحو: S038_T2_M_Pre_NNAS_W_C. وهي بالترتيب من اليسار (مفصولةٌ بشرطةٍ سفليةٍ): رقم الطالب، رقم النص، نوع الطالب، المرحلة العامة، ناطق بالعربية كونها لغته الأم أم ناطق بغيرها، نوع النص (مكتوبٌ أم منطوقٌ)، مكان تحرير أو تسجيل النص (Alfaifi & Atwell, 2014). ويبين الشكل رقم (١٠) الصفحة الرئيسة للمدونة اللغوية لمتعلمي اللغة العربية.



الشكل رقم (١٠) الصفحة الرئيسة للمدونة اللغوية لمتعلمي اللغة العربية

المدونة العربية القرآنية The Quranic Arabic Corpus

هي مدونةٌ لغويةٌ مُحَشّوَّةٌ annotated (corpus.quran.com) توضح اللغة العربية في ضوء الأجرومية، والنحو، والصرف لكل كلمةٍ من كلمات القرآن الكريم (Dukes & Habash, 2010; The Quranic Arabic Corpus). وتوفر هذه المدونة اللغوية ثلاثة مستوياتٍ من التحليل، هي:

١. التحشية الصرفية morphology annotation:

توجد في هذه المدونة اللغوية في قسم فرعي داخل الموقع باسم «كلمةٌ كلمةٌ Word by Word». وهي في حقيقة الأمر ليست تحشيةً صرفيةً فقط، بل هي أيضًا تحشيةٌ نحويةٌ لكل كلمةٍ من كلمات القرآن الكريم، مسبوقةٌ بنقحرتها بالحروف اللاتينية، ومتبوعةٌ بترجمةٍ لها إلى الإنجليزية.

ويمكن البحث في واجهة الاستخدام هنا وفقًا للسور والآيات. ويوضح الشكل رقم (١١) التحشية الصرفية النحوية في المدونة العربية القرآنية.

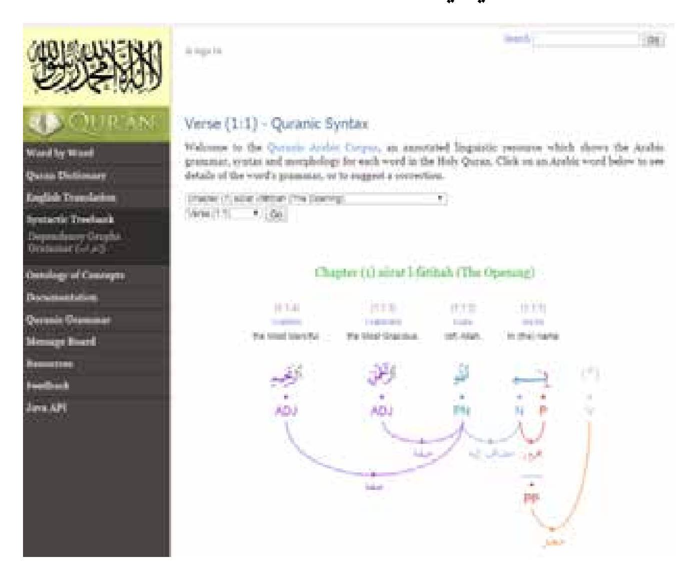


الشكل رقم (١١) التحشية الصرفية النحوية في المدونة العربية القرآنية

٢. بنك الأشجار النحوي **syntactic treebank**:

موجودةٌ أيضًا في قسمٍ فرعي داخل الموقع باسم «Quranic Syntax». ويتم هنا عرض كل كلمةٍ وفقًا لموقعها الإعرابي داخل الجملة بشكلٍ رسوميٍ، مع إمكانية تقديم إعرابٍ كاملٍ لكل جملةٍ.

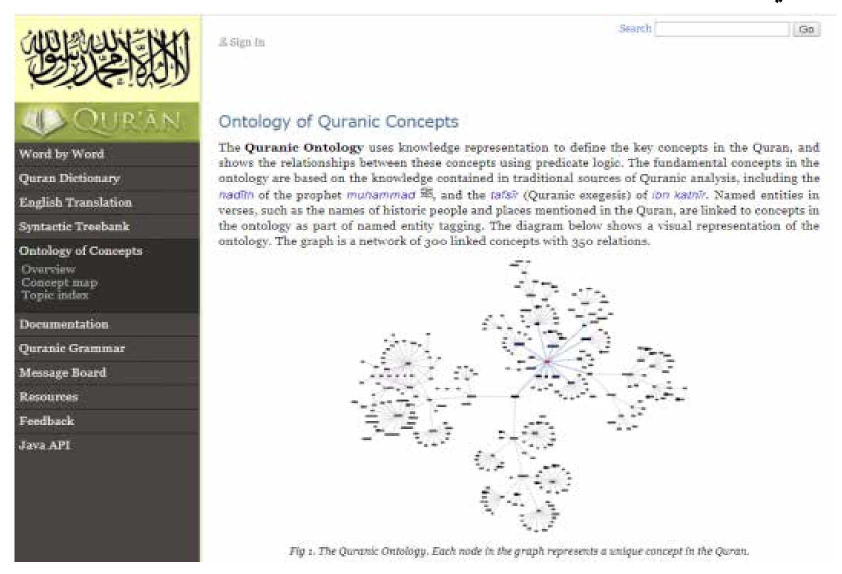
ويمكن البحث في واجهة الاستخدام هنا وفقًا للسور والآيات أيضًا. ويبين الشكل رقم (١٢) بنك الأشجار النحوي في المدونة العربية القرآنية.



الشكل رقم (١٢) بنك الأشجار النحوي في المدونة العربية القرآنية

٣. الأنطولوجيا الدلالية semantic ontology:

موجودةٌ في قسمٍ فرعيٍ بالموقع باسم «أنطولوجيا المفاهيم القرآنية Ontology of Quranic Concepts». حيث يتم استخدام تمثيل المعرفة للتعريف بمفاهيم الأساس في القرآن الكريم في شكل علاقاتٍ منطقيةٍ. ويوضح الشكل رقم (١٣) الأنطولوجيا الدلالية في المدونة العربية القرآنية.



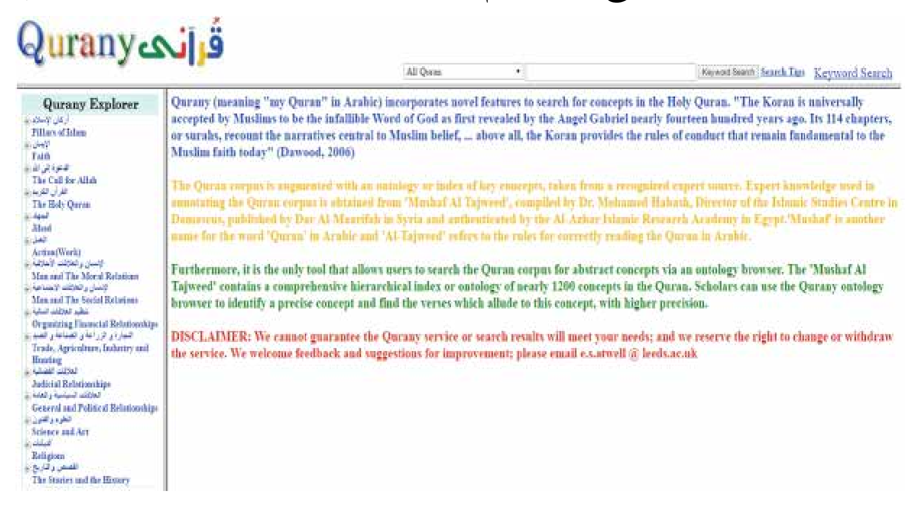
الشكل رقم (١٣) الأنطولوجيا الدلالية في المدونة العربية القرآنية

مدونة قرآني Qurany

هي مدونةٌ لغويةٌ لآي القرآن الكريم (quranytopics.appspot.com). وقد تم تصنيف الآيات فيها وفقًا للمفاهيم أو الموضوعات المفتاحية المعتمِدة على مصحف التجويد للدكتور محمد حبش، والتي تم توثيقها بواسطة الأزهر الشريف. إذ تُمكِّن هذه المدونة اللغوية المستفيدين من البحث عن المفاهيم الواردة بآيات القرآن وفقًا لأنطولوجيا هرميةٍ سهلة التصفح تم بناؤها باستخدام لغة الترميز القابلة للتمديد XML.

وقد تم إحصاء المفاهيم الواردة بهذه المدونة اللغوية في ١٢٠٠ مصطلحٍ. وبالبحث بأحدها فإنه يمكن استرجاع كافة الآيات التي تدل على المصطلح المبحوث به. ويتم عرض كل مصطلحٍ في سياقه (تكشيف نصوص).

وعلاوةً على ذلك فإن هذه المدونة اللغوية تتيح إمكانية البحث بالكلمات المفتاحية وعرض النتائج كما وردت في سياقاتها القرآنية. ويتم عرض النتائج المسترجعة مصحوبةً بترجمةٍ إلى الإنجليزية. ويوضح الشكل رقم (١٤) الصفحة الرئيسة للمدونة اللغوية «قرآني».



الشكل رقم (١٤) الصفحة الرئيسة للمدونة اللغوية "قرآني"

استقصاء المدونات اللغوية العربية Querying Arabic Corpora

هي مدونةٌ لغويةٌ تابعةٌ لجامعة ليدز بإنجلترا (corpus.leeds.ac.uk /query-ar.html) تتيح البحث في ست مدوناتٍ لغويةٍ فرعيةٍ، هي:

١. المدونة اللغوية للإنترنت.
٢. المدونة اللغوية لجريدة الحياة (من عام ١٩٩٩ وحتى عام ٢٠٠١).
٣. المدونة اللغوية لـ ويكيبيديا.
٤. المدونة اللغوية للعربية المعاصرة.
٥. المدونة اللغوية القانونية العربية.
٦. المدونة اللغوية العربية لعلوم الحاسب.

ويبين الشكل رقم (١٥) الصفحة الرئيسة لمدونة استقصاء المدونات اللغوية العربية.

Querying Arabic Corpora

Arabic Internet Al Hayat News Arabic Wikipedia Corpus of Contemporary Arabic Computer Science corpus
Arabic legal texts, v2
Submit Reset CQP syntax only (Examples) Click here for getting help on the query interface

Set parameters of your query

Concordance

Context: 60c (c for characters, w for words)

Sort by: Document Frequency lemma word
Then by: left right

Output: 100 lines

Collocations

Collocation scores: Mutual Information Dice T-score Loglikelihood score

Context: 0 words on the left 1 words on the right

POS tag of the collocate POS tags

I also created the lists of the most frequent word forms in Internet, LDC, Wikipedia and CCA, as well as in the legal corpus.
After lemmatisation done by Majdi Sawalha there is also the frequency list of lemmas and rootsin the Arabic Internet corpus.

The corpora are:

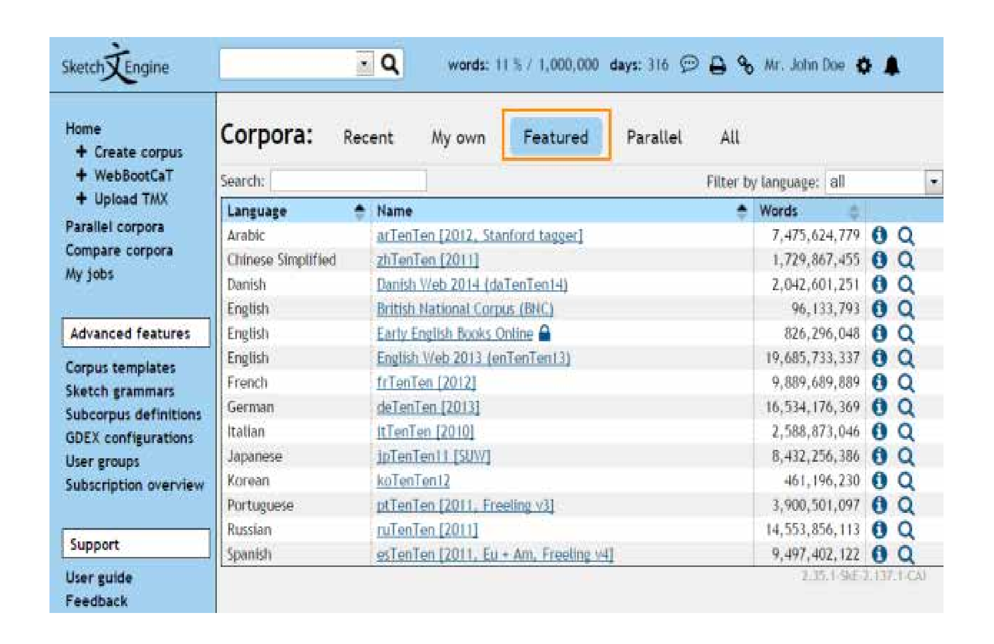
1. The Internet corpus was compiled using the procedure described in my paper in the WaCky book.
2. The Al Hayat corpus — from Al Hayat data (1999-2001) compiled by the LDC.
3. The Wikipedia corpus — from the public Wiki data retrieved on July 28, 2008.
4. CCA corpus — from Latifa Al-Sulaiti
5. The Arabic Legal Corpus — from keywords collected by Hatem El-Farahaty, a Leeds PhD student.
6. Computer Science corpus of Arabic — from keywords collected by Latifa Al-Sulaiti

The interface was developed by Serge Sharoff, contact me at s.sharoff@leeds.ac.uk, if you have further queries.

الشكل رقم (١٥) الصفحة الرئيسة لمدونة استقصاء المدونات اللغوية العربية

سكتش إنجين Sketch Engine

أُسست هذه المدونة اللغوية المعتمدة على الشبكة العنكبوتية على يد العالم آدم كيلجاريف في عام ٢٠٠٣ (www.sketchengine.co.uk). وتتيح هذه المدونة التعامل مع ٤٠٠ مدونةٍ لغويةٍ فرعيةٍ لأكثر من ٩٠ لغةً، من بينها اللغة العربية. ويصل حجم بعض هذه المدونات اللغوية إلى ٢٠ مليار كلمةٍ. وإضافة إلى إمكانات البحث التي تقدمها سكتش إنجين، فإنها توفر إيضًا إمكانية رفع المدونات اللغوية التي ينشئها المستفيدون والباحثون بأنفسهم، أو من خلال تجميع النصوص من الشبكة العنكبوتية. ويوضح الشكل رقم (١٦) واجهة الاستخدام الرئيسة لسكتش إنجين.



Language	Name	Words
Arabic	arTenTen [2012, Stanford tagger]	7,475,624,779
Chinese Simplified	zhTenTen [2011]	1,729,867,455
Danish	Danish Web 2014 (daTenTen14)	2,042,601,251
English	British National Corpus (BNC)	96,133,793
English	Early English Books Online	826,296,048
English	English Web 2013 (enTenTen13)	19,685,733,337
French	frTenTen [2012]	9,889,689,889
German	deTenTen [2013]	16,534,176,369
Italian	itTenTen [2010]	2,588,873,046
Japanese	jpTenTen11 [SUW]	8,432,256,386
Korean	koTenTen12	461,196,230
Portuguese	ptTenTen [2011, Freeling v3]	3,900,501,097
Russian	ruTenTen [2011]	14,553,856,113
Spanish	esTenTen [2011, Eu + Am, Freeling v4]	9,497,402,122

الشكل رقم (١٦) واجهة الاستخدام الرئيسة لسكتش إنجين

وتتوافر هذه المدونة اللغوية على مستوياتٍ عدةٍ من التحليل، منها:

مخطط الكلمات Word Sketch

تتيح إمكانية التحليل هذه ملخصًا لسلوك الكلمات يتضمن عرض المتلازمات اللغوية للكلمات، مصنفًا وفق العلاقات النحوية؛ مثل: الكلمات التي تعمل بصفتها فاعلًا، والكلمات التي تعمل بصفتها مفعولًا، وهكذا...

ويتفرع من هذا التحليل تحليلان آخران مترتبان عليه؛ يتيح الأول إمكانية عقد مقارنةٍ بين كلمتين في لغةٍ واحدةٍ، فيما يتيح الثاني المقارنة بين كلمةٍ في لغةٍ ما ومقابلها في لغةٍ آخرى، وذلك في إطار مخطط الكلمات. ويبين الشكل رقم (١٧) إمكانات التحليل التي يقدمها مخطط الكلمات في سكتش إنجين لكلمة «team».

team (noun) Alternative PoS: verb (478)
British National Corpus (BNC) freq = 22,482 (200.21 per million)

modifiers of "team"	13,919	0.62
management +	422	9.21
football +	207	8.63
project +	188	8.35
england +	142	8.05
research +	164	7.83
rescue	58	7.76
display	51	7.60
cup	55	7.45
design	87	7.38

nouns and verbs modified by "team"	3,166	0.14
spirit +	112	9.15
mate	53	8.75
leader +	133	8.26
coach	40	8.09
manager +	131	8.05
member +	197	8.01
effort	72	7.94
championship	49	7.77
selection	38	7.73

verbs with "team" as object	4,616	0.21
lead +	205	8.48
head	82	8.26
join +	113	8.04
pick	47	7.79
field	26	7.43
assemble	25	7.17
beat	34	7.01
negotiate	26	7.00
captain	18	6.92
send	33	6.86
strengthen	22	6.79
investigate	27	6.77
select	27	6.74
visit	36	6.53

verbs with "team" as subject	6,200	0.28
win	56	7.97
play +	105	7.86
work +	109	7.53
lose	40	6.78
consist	31	6.78
perform	27	6.74
compete	22	6.70
find	57	6.55
comprise	21	6.46
prepare	22	6.45
take +	100	6.36

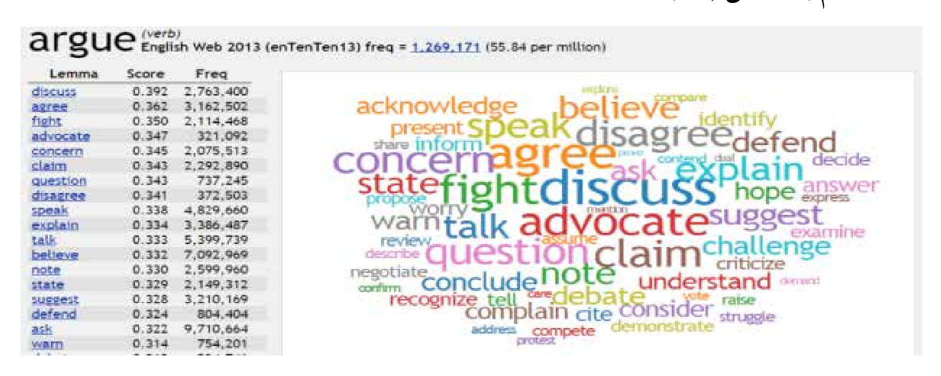
"team" and/or ...	2,244	0.10
football	12	7.15
cast	8	6.75
search	9	6.71
group	21	6.55
squad	7	6.55
individual	12	6.41
husband	12	6.37
player	10	6.35
supporter	7	6.19
afternoon	7	6.17
fan	6	6.11
panel	6	6.11
specialist	6	6.08
tale	10	6.07
member	18	6.01
department	10	5.93
management	12	5.91
manager	12	5.88

الشكل رقم (١٧) إمكانات التحليل التي يقدمها مخطط الكلمات في سكتش إنجين لكلمة ”team“

المكنز thesaurus

تُمكِّن هذه الأداة المستفيدين والباحثين من توليد مرادفات الكلمات اعتمادًا على خوارزميات تضطلع بالبحث عن الكلمات التي ترد في سياقاتٍ متشابهةٍ في إحدى المدونات اللغوية. وفي الوقت نفسه، تتيح هذه الأداة المقارنة بين الكلمات، والسياقات الفعلية التي وردت بها هذه الكلمات، مع إبراز الكلمات ذات السلوك التلازمي المشابه ولكنها تبدو بعيدة الصلة عن الكلمة المعنية.

ويتم عرض المرادفات وفقًا للأسرة اللغوية للكلمة، وعدد مرات تكرارها، ونسبتها. إضافةً إلى عرضها في شكل خريطةٍ سحابيةٍ. ويوضح الشكل رقم (١٨) صفحة المكنز المستخدم بسكتش إنجين لكلمة «argue».



argue (verb) English Web 2013 (enTenTen13) freq = 1,269,171 (55.84 per million)

Lemma	Score	Freq
discuss	0.392	2,763,400
agree	0.362	3,162,502
fight	0.350	2,114,468
advocate	0.347	321,092
concern	0.345	2,075,513
claim	0.343	2,292,890
question	0.343	737,245
disagree	0.341	372,503
speak	0.338	4,829,660
explain	0.334	3,386,487
talk	0.333	5,399,739
believe	0.332	7,092,969
note	0.330	2,599,960
state	0.329	2,149,312
suggest	0.328	3,210,169
defend	0.324	804,404
ask	0.322	9,710,664
warn	0.314	754,201

الشكل رقم (١٨) صفحة المكنز المستخدم بسكتش إنجين لكلمة “argue”

تكشيف النصوص concordance

يتوافر هذا التحليل على تقديم قائمة بكافة السياقات التي تضم الكلمة أو العبارة المبحوث بها في المدونة اللغوية، مصحوبةً ببعض الكلمات عن يمينها، وبعض الكلمات عن يسارها. ويتم إبراز هذه الكلمة أو العبارة بلونٍ مختلفٍ في منتصف صفحة العرض لتمييزها. يوضح الشكل رقم (١٩) آلية عرض كشاف النصوص في سكتش إنجين.



الشكل رقم (١٩) آلية عرض كشاف النصوص في سكتش إنجين

وينبثق من هذا التحليل تحليلٌ آخر يتيح إمكانية استرجاع سياقات إحدى الكلمات في لغةٍ معينةٍ، وبجوارها مقابلاتها في سياقاتها المختلفة في لغةٍ أخرى. الأمر الذي يتيح مرونةً أكبر أثناء المقارنة بين المفردات واستخداماتها في اللغات المختلفة.

قائمة الكلمات Word List

تقوم سكتش إنجين بعمل فرزٍ للكلمات وفقًا لتكرار ترددها من خلال هذا التحليل. ويبين الشكل رقم (٢٠) آلية عرض الكلمات وفقًا لتكرار ترددها بسكتش إنجين.

word (n-grams)	frequency
صلى الله عليه وسلم	58,911
الله صلى الله عليه	20,156
رسول الله صلى الله	19,074
النبي صلى الله عليه	17,887
صلى الله عليه وآله	14,697
الله عليه وآله وسلم	6,797
بسم الله الرحمن الرحيم	5,644
السلام عليكم ورحمة الله	4,945
عليكم ورحمة الله وبركاته	4,716
الرسول صلى الله عليه	4,023
صلى الله عليه و	3,758
محمد صلى الله عليه	3,089
الله عليه و سلم	2,994
علي بن أبي طالب	2,857

الشكل رقم (٢٠) آلية عرض الكلمات وفقًا لتكرار ترددها بسكتش إنجين

استخراج المصطلحات Term Extraction

تدعم سكتش إنجين إمكانية استخراج المصطلحات والكلمات المفتاحية من المدونات اللغوية مستخدمةً مجموعةً من المعايير اللغوية لتحديد هذه المصطلحات. حيث يمكن للباحثين أن يقوموا برفع مدوناتهم اللغوية على سكتش إنجين التي تقوم بدورها بتحليلٍ لهذه المدونات اللغوية مستخرجةً المصطلحات والكلمات المفتاحية التي تعبر عنها. ويوضح الشكل رقم (٢١) شاشة عرض المصطلحات والكلمات المفتاحية وفق تحليل سكتش إنجين.

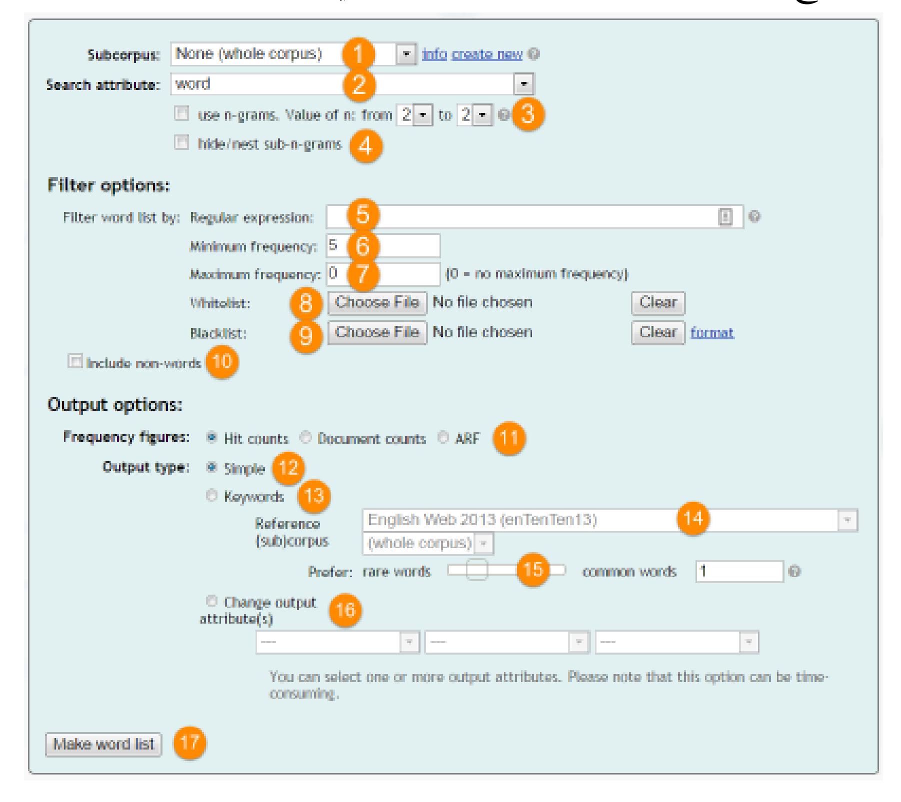
Keywords	Score	F	RefF
co2	176.40	12,394	0
biodiversity	34.38	14,663	65,693
ecosystems	32.71	11,893	54,163
emissions	31.24	54,232	306,028
unep	30.96	3,231	6,603
watershed	28.70	10,555	54,953
deforestation	28.06	5,200	21,498
climate	27.56	116,973	786,522
biomass	26.71	9,341	51,698
habitats	26.63	9,606	53,974
wetlands	26.47	9,030	50,123
greenhouse	26.14	22,514	145,622
desertification	25.45	2,448	5,194
wwf	25.25	3,678	16,457
dioxide	24.76	15,611	103,261
renewable	24.24	31,170	223,599
redd	23.86	2,938	10,169
wetland	23.63	5,226	28,175
ghg	23.52	3,930	18,255
carbon	23.37	65,299	500,429
conservation	23.14	42,635	325,757
stormwater	22.93	4,148	20,726
ecological	22.59	14,692	106,963
pollution	22.47	27,926	215,719

Terms	Score	F	RefF
climate change	39.64	54,341	238,935
greenhouse gas	32.65	11,431	51,682
water quality	29.19	9,823	49,251
carbon dioxide	26.07	13,115	79,874
renewable energy	24.73	16,926	113,194
sea ice	22.66	2,824	10,489
global warming	22.15	17,102	129,357
global climate	22.11	3,467	16,403
fossil fuel	20.77	4,052	23,470
sustainable development	20.64	6,099	41,897
clean energy	19.57	4,694	31,732
air pollution	17.53	3,941	29,033
water management	16.23	2,222	12,962
land use	15.98	4,729	42,162
low carbon	15.76	2,137	12,751
human health	15.69	3,416	27,817
organic matter	15.68	2,364	15,539
coal-fired power	15.23	1,654	7,819
global climate change	14.90	1,681	8,622
solar energy	14.71	6,211	65,411
energy efficiency	14.59	6,798	73,286
environmental impact	14.47	3,649	34,236
wind power	14.43	3,525	32,762
water use	14.40	2,152	15,495

الشكل رقم (٢١) شاشة عرض المصطلحات والكلمات المفتاحية وفق تحليل سكتش إنجين

تحليل المتتابعات اللفظية n-grams

هو تسلسلٌ لعددٍ من الأرقام، أو الكلمات، أو الحروف... إلخ. غير إنه يشير عادةً إلى تسلسلٍ من الكلمات من وجهة نظر المدونات اللغوية. ويُستخدم مصطلح «unigram» للإشارة إلى كلمةٍ واحدةٍ، ويشير مصطلح «bigram» إلى تسلسلٍ من كلمتين، فيما يشير مصطلح «trigram» إلى تسلسلٍ من ثلاث كلماتٍ، وهكذا. والرابط الوحيد الذي يجمع بين الكلمات في تحليل المتتابعات اللفظية هو ظهورها بجوار بعضها البعض؛ ولذا ليس بالضرورة أن تكون هناك علاقةٌ واضحةٌ بين هذه الكلمات. الأمر الذي يعني أن مخرجات تحليل المتتابعات اللفظية ليست جميعها متلازماتٍ لغويةً collocates، غير إنه يسهم في استخراجها. ويوضح الشكل رقم (٢٢) كيفية ضبط إعدادات تحليل المتتابعات اللفظية لاستخراج تسلسلات الكلمات والمتلازمات اللغوية في سكتش إنجين.



الشكل رقم (٢٢) كيفية ضبط إعدادات تحليل المتتابعات اللفظية لاستخراج تسلسلات الكلمات والمتلازمات اللغوية في سكتش إنجين

المدونة اللغوية للإنجليزية الأمريكية المعاصرة
The Corpus of Contemporary American English (COCA)

تعرف اختصارًا باسم «كوكا». وأنشئت على يد مارك ديفيس Mark Davies الباحث بجامعة بريجهام يانج الأمريكية Brigham Young University (corpus.byu.edu /coca). وتُعد هذه المدونة واحدةٌ من أكبر المدونات اللغوية الإنجليزية المتاحة مجانًا، وأكثرها استخدامًا. وترتبط بها الكثير من المدونات اللغوية الإنجليزية الأخرى. وتحوي المدونة أكثر من ٥٣٣ مليون كلمة؛ بمعدل إضافة ٢٠ مليون كلمة جديدة كل عام. وقد تم تجميع نصوص هذه المدونة فيما بين عامي ١٩٩٠ و ٢٠١٥. وتضم خمسة أنواعٍ أدبيةٍ؛ هي: النصوص المنطوقة، والقصص، والمجلات، والصحف، والدوريات الأكاديمية. ويوضح الشكل رقم (٢٣) الصفحة الرئيسة لمدونة كوكا.

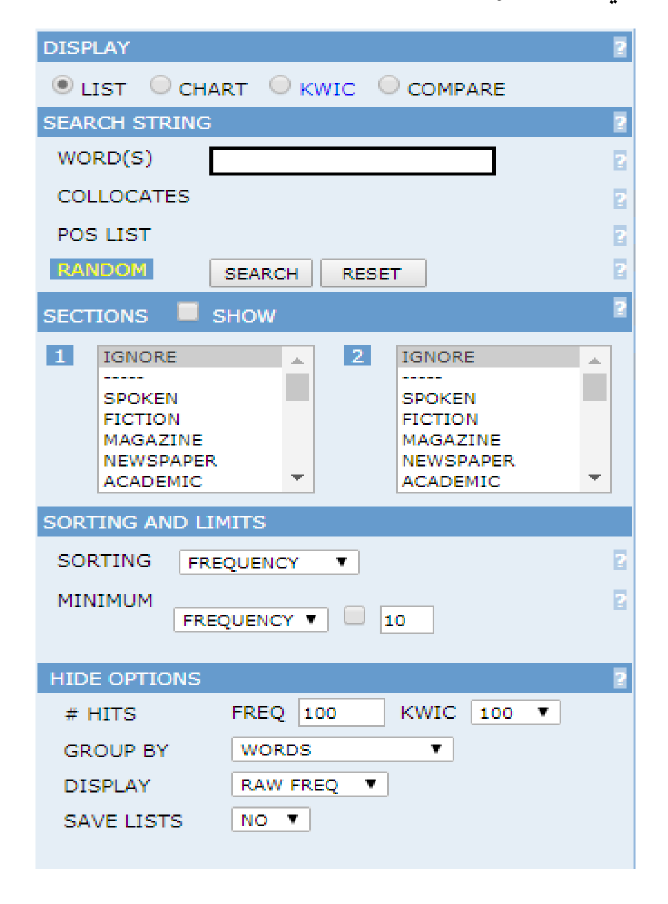


الشكل رقم (٢٣) الصفحة الرئيسة لمدونة كوكا

وتتيح الصفحة الرئيسة لمدونة كوكا إمكانية البحث العام للكلمات من خلال تبويبة «البحث Search»، ومن ثم استرجاع مدى تكرار ترددها من خلال تبويبة «التردد Frequency»، وعرض السياقات من خلال تبويبة «السياق Context».

تُمَكِّن تبويبة البحث الاسترجاعَ بالكلمة، أو بالعبارة، أو بجزءٍ من الكلمة، أو بالأسرة اللغوية، أو بقسم الكلمة، أو باثنين أو أكثر من إمكانات الاسترجاع هذه. إضافةً إلى إمكانية استرجاع المتلازمات اللغوية في حدود عشر كلماتٍ يمينًا ويسارًا.

ومن الممكن أن يتم تقييد البحث في مدونة كوكا بمدى تكرار التردد، ومقارنة تكرار تردد الكلمات، والعبارات، والأبنية النحوية، وذلك وفق النوع الأدبي، أو التحديد الزمني. كما أن مدونة كوكا تتيح إمكانية البحث الدلالي في نصوصها للأضداد والمترادفات، والمقارنة بينها في الأنواع الأدبية الخمسة التي تضمها. ويوضح الشكل رقم (٢٤) صفحة البحث في مدونة كوكا.



الشكل رقم (٢٤) صفحة البحث في مدونة كوكا

وينبثق من مدونة كوكا الكثير من المدونات اللغوية الأخرى، منها: المدونة اللغوية الوطنية البريطانية British National Corpus (BNC)، والمدونة اللغوية للأخبار على الشبكة العنكبوتية (ناو) News on the Web (NOW Corpus)، والمدونة اللغوية لكتب جوجل Google Books Corpus، وغيرها.

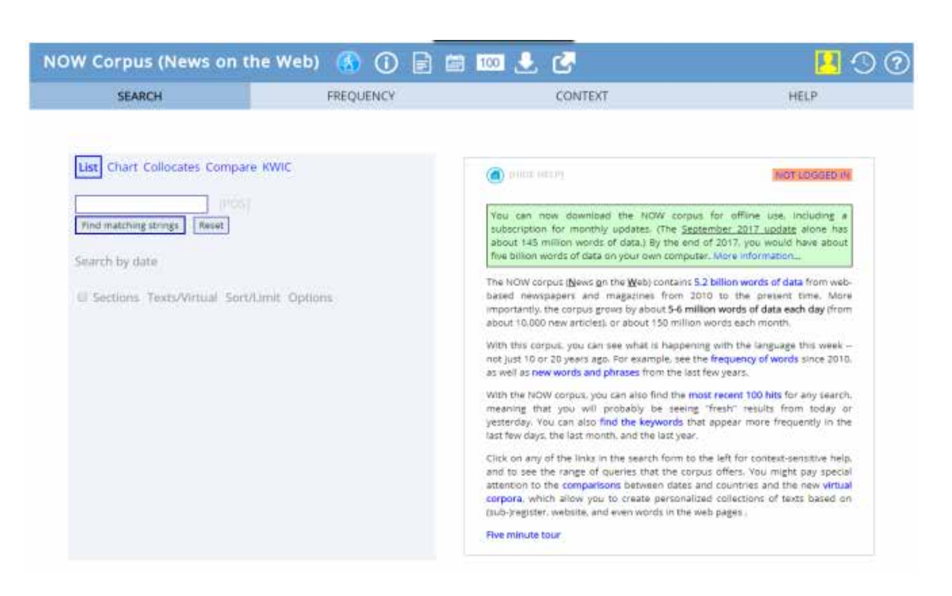
المدونة اللغوية للأخبار على الشبكة العنكبوتية

News on the Web (NOW Corpus)

تضم هذه المدونة اللغوية حوالي ٣ مليار كلمةٍ للأخبار الإنجليزية المتاحة على الشبكة العنكبوتية في الفترة من عام ٢٠١٠ وحتى الآن. وتنمو هذه المدونة اللغوية بمعدلٍ يبلغ ٤ مليون كلمةٍ يوميًا مأخوذةً من حوالي ١٠٠٠٠ مقالٍ إخباري (corpus.byu.edu /now).

وتقوم فكرة هذه المدونة اللغوية على إمكانية التحقق من الموضوعات المفتاحية الجارية على الساحة أولًا بأولٍ. فعلى سبيل المثال، يُمْكن للمستفيدين استرجاع الكلمات وفقًا لتكرار ترددها منذ عام ٢٠١٠. كما يمكنهم أيضًا التحقق من المصطلحات والعبارات الجديدة التي تعكس ما يُستجَد من موضوعاتٍ، وكذلك الكلمات المفتاحية الدالة على أحدث الأخبار.

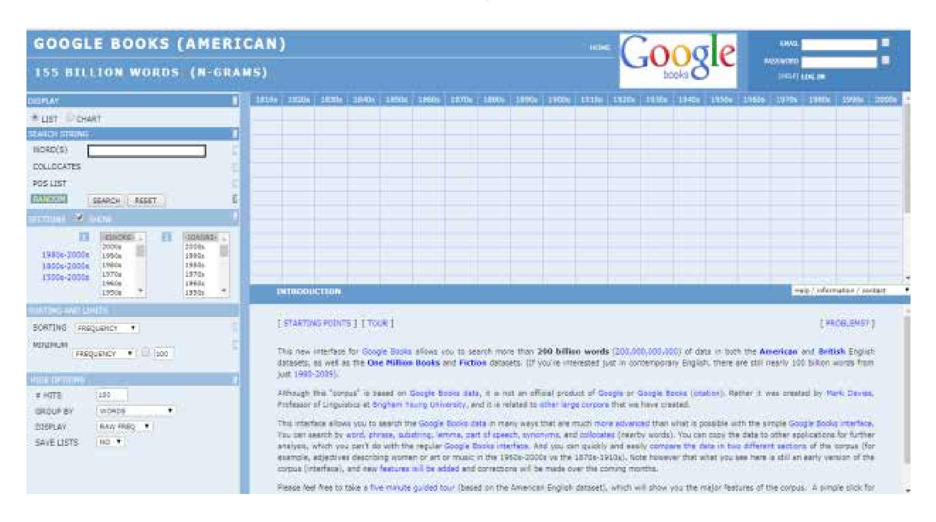
وإضافةً إلى ذلك فإن المدونة اللغوية «ناو» تتيح إمكانية عقد مقارناتٍ بين الفترات الزمنية، والدول، والمواقع الإخبارية. ويوضح الشكل رقم (٢٥) الصفحة الرئيسة للمدونة اللغوية «ناو».



الشكل رقم (٢٥) الصفحة الرئيسة للمدونة اللغوية "ناو"

المدونة اللغوية لكتب جوجل Google Books Corpus

تأتي هذه المدونة اللغوية مشروعًا تقدم به مارك ديفيس Mark Davies الباحث بجامعة بريجهام يانج الأمريكية Brigham Young University (googlebooks.byu.edu). وتضم هذه المدونة اللغوية في حقيقة أمرها كل ما تحويه قاعدة بيانات كتب جوجل Google Books التي تغطي ملايين الكتب الإنجليزية (أكثر من ٢٠٠ مليار كلمةٍ إنجليزيةٍ) خلال الخمسة قرونٍ المنصرمة (الفترة من عام ١٥٠٠ وحتى عام ٢٠٠٩). ويعرض الشكل رقم (٢٦) الصفحة الرئيسة لمدونة كتب جوجل.



الشكل رقم (٢٦) الصفحة الرئيسة لمدونة كتب جوجل

حيث قام ديفيس بإخضاع هذه المدونة اللغوية للتحليل من خلال واجهةٍ تعاملٍ متقدمةٍ تُمكِّن من البحث فيها على نحوٍ أكثر عمقًا منه في واجهة التعامل البسيطة. وتتيح واجهة التعامل المتقدمة هذه تحليلًا كميًا على مستوى الكلمة، أو العبارة، أو الجملة، أو الجذر، أو المترادفات، أو أقسام الكلمات، أو المصطلحات المتلازمة. فضلًا عن أنها تتيح إمكانية نسخ البيانات في حالة أن أراد المستفيدون القيام بتحليلاتٍ أخرى غير متاحةٍ من خلال واجهة التعامل المتقدمة. وكذا، يتمكن المستفيدون من التحقق من الكلمات المفتاحية الدالة على موضوعٍ معينٍ خلال فترةٍ زمنيةٍ محددةٍ. وفي الوقت نفسه يُمْكنهم إجراء مقارناتٍ بين الكلماتٍ المفتاحية الدالة على موضوعين أو أكثر خلال الفترات الزمنية المختلفة (Davies, 2011). ويتم عرض النتائج في شكل خريطةٍ رسوميةٍ توضح التوزيعات الإحصائية لما تم البحث به.

برمجيات معالجة وتحليل المدونات اللغوية العربية

توجد بعض البرمجيات التي تدعم معالجة المدونات اللغوية العربية، ولعل من أشهرها برنامج أدوات وورد سميث WordSmith Tools، وبرنامج أدوات معالجة المدونات اللغوية العربية The Arabic Corpus Processing Tools (ACPTs)، وبرنامج آنتكونك AntConc (Anthony, 2017)، وسكيتش إنجين (Sketch Engine 2015)، وبرنامج استفسارات مدونة إنتليتيكست IntelliText Corpus Queries (Wilson et al., 2010; Sharoff, 2014)، ومعالج استفسارات المدونات اللغوية Corpus Query Processor (CQP. وسوف نكتفي هنا بتناول أول برنامجين فقط بقدرٍ من التفصيل.

برنامج أدوات وورد سميث WordSmith Tools

هو حزمة برمجياتٍ مدفوعة الأجر تُستخدم في تحليل النصوص المتاحة بلغاتٍ عدةٍ، ومنها النصوص العربية. ويقوم على هذا البرنامج مايك سكوت Mike Scott الباحث بجامعة ليفربول University of Liverpool وبالتعاون من جامعة أكسفورد Oxford University (Scott, 2016). ويوضح الشكل رقم (٢٧) واجهة الاستخدام الرئيسة لبرنامج أدوات وورد سميث.



الشكل رقم (٢٧) واجهة الاستخدام الرئيسة لبرنامج أدوات وورد سميث

العمليات الفنية في برنامج أدوات وورد سميث

يتوافر برنامج وورد سميث على تقديم ثلاث عملياتٍ فنيةٍ رئيسةٍ متمثلةٍ في ثلاثة نظمٍ فرعيةٍ، هي:

١. كشاف الكلمات المفتاحية في السياق «كونكورد Concord»:

يُستخدَم هذا النظام الفرعي في إنشاء كشافٍ بالكلمات المفتاحية في السياق، وذلك من خلال البحث بكلمةٍ معينةٍ في مدونةٍ لغويةٍ مُعدَّةٍ مسبقًا.

٢. قائمة الكلمات «وورد لِست WordList»:

يرتِب هذا النظام الفرعي الكلمات أو أشكالًا محددةً من الكلمات المتضمَنة بالمدونة اللغوية وفقًا لبياناتها الإحصائية وتكرار ترددها.

٣. مُحلل الكلمات المفتاحية «كي وورد KeyWord»:

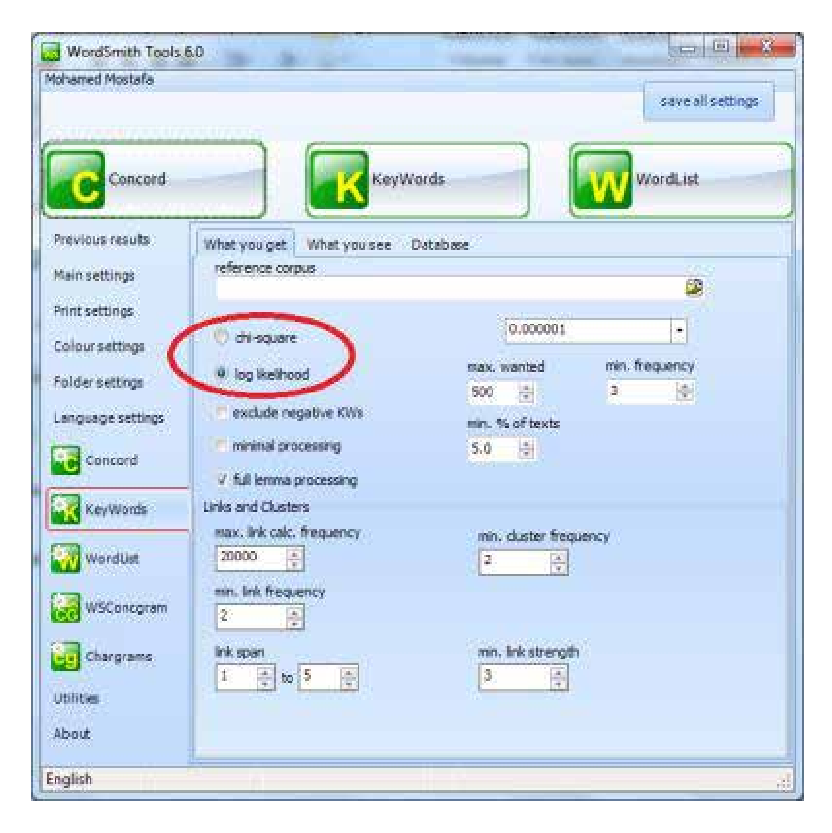
يُستفاد من هذا النظام الفرعي في إنشاء قوائم بالكلمات المفتاحية للمدونات اللغوية وفقًا لمجموعةٍ من المعايير الإحصائية، ويرتبها تبعًا لدلالتها الإحصائية.

ويُقصد بالكلمات المفتاحية في هذا النظام الفرعي تلك الكلمات ذات التردد العالي غير العادي وفقًا لبعض المعدلات الإحصائية. إذ يقوم هذا النظام الفرعي بعقد مقارنةٍ إحصائيةٍ بين قائمتي كلماتٍ تم إعدادهما مسبقًا باستخدام النظام الفرعي الخاص بقائمة الكلمات «وورد لِيست WordList». فالقائمة الأولى (المدونة اللغوية موضوع التحليل والدراسة، وهي الأقل حجمًا) هي تلك التي سيتم تكشيف كلماتها.

أما القائمة الثانية، الأكبر حجمًا، فهي تعمل ملفًا مرجعيًا أو مدونةً مرجعيةً reference corpus (ذ م RC) يتم في مقابلها عقد هذه المقارنة (Scott, 1997).

ويتيح برنامج وورد سميث إمكانية الاختيار بين طريقتين إحصائيتين لتحليل الكلمات المفتاحية، هما طريقة اختبار مربع كاي chi-square للدلالة الإحصائية، وطريقة احتمالات سجل الأداء log likelihood التي تعطي تقديراتٍ أفضل حول نسبة المفتاحية keyness في النصوص (Dunning, 1993). ويوضح الشكل رقم (٢٨) كيفية الاختيار بين هاتين الطريقتين الإحصائيتين في برنامج أدوات وورد سميث.

وللمزيد حول آلية عمل تكشيف الكلمات المفتاحية في المدونات اللغوية، يمكن الرجوع إلى الجزئية الخاصة بعمليات معالجة البيانات على مستوى الإخراج، وتكشيف الكلمات ضمن الفصل الثالث.



الشكل رقم (٢٨) طريقة الاختيار بين اختبار مربع كاي للدلالة الإحصائية، وطريقة احتمالات سجل الأداء في برنامج أدوات وورد سميث

وإضافةً إلى هذه النظم الفرعية الثلاثة، فإن برنامج وورد سميث يتيح مجموعةً أخرى من الإمكانات التي تضفي قدرًا أدق من التحليل على المدونات اللغوية، منها:

١. تشتت الكلمات المفتاحية keyword dispersion: يتم فيه التحقق إحصائيًا من الدرجة التي تتوزع بمقتضاها مجموعةٌ من القيم بشكلٍ موحدٍ ضمن مجتمع الدراسة. وتتراوح درجة التشتت في برنامج وورد سميث ما بين ٠ و ١. فكلما اقتربت القيمة من ١ ازداد التشتت. فيما تُعبر القيمة ٠ عن تماسكٍ شديدٍ في التشتت burstiness (Katz, 1996; Scott, 2016).

٢. تحليل التلازم اللغوي collocation: أو ما يُطلَق عليه أحيانًا استرجاع المتلازمات اللغوية؛ أي التحقق من الكلمات التي تتصاحب أو تتلازم بكثرةٍ مع الكلمات المفتاحية بالمدونة اللغوية.

٣. الكشف عن القوالب اللغوية patterns: أو ما يُطلَق عليه أحيانًا تحليل الأسلوب العِباري للكلمات phraseology؛ أي التحقق من البيئة اللغوية للكلمات المفتاحية، أو الكلمات بصفةٍ عامةٍ، داخل العبارات (Scott, 2016).

برنامج أدوات معالجة المدونات اللغوية العربية

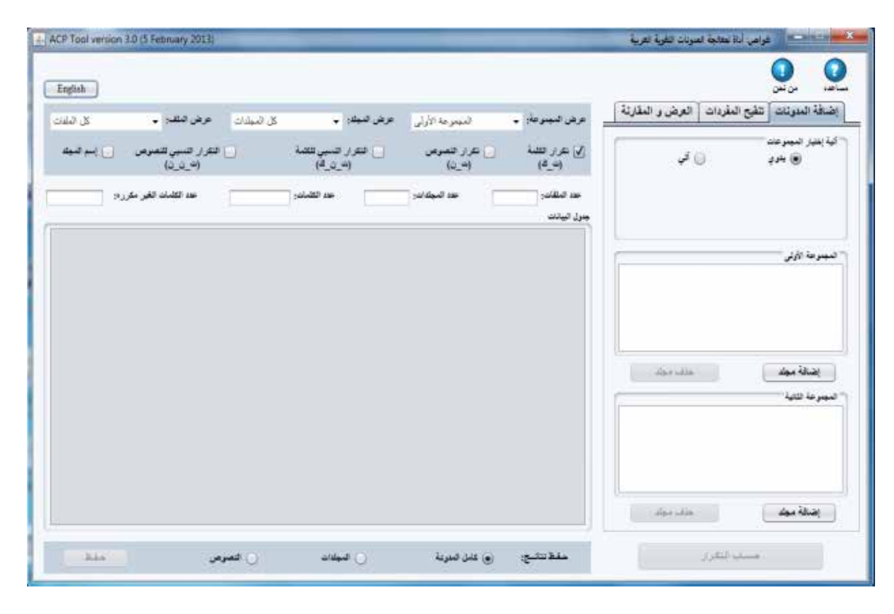
The Arabic Corpus Processing Tools (ACPTs)

كان يُعرف في السابق باسم «غَوَّاص Ghawwas»،[١] وكذلك كان يُعرف باسم «خَوَاص Khawas». ويقوم عليه عبد المحسن الثبيتي بمساعدة آخرين من مدينة الملك عبد العزيز للعلوم والتقنية (Almujaiwel & Al-Thubaity, 2016; Al-Thubaity, Khan, Al-Mazrua & Al-Mousa, 2013).

وهو برنامجٌ مجانيٌ مفتوح المصدر يتسم بإمكاناته الكبيرة في معالجة النصوص العربية. والبرنامج ذو واجهتي استخدامٍ؛ عربيةٌ، وإنجليزيةٌ. ويمكن للبرنامج معالجة النصوص الأخرى غير العربية، كالإنجليزية، والفرنسية. ويستطيع البرنامج عرض تكرار تردد هياكل الكلمات tokens، والكلمات الفريدة types، والوثائق. ويدعم البرنامج أشكالًا مختلفةً من الملفات؛ النصوص البسيطة txt، وملفات الوورد doc, docx، وصفحات إتش تي إم إل html. كما يدعم البرنامج نظامي التشفير آنسي ANSI، و يو تي إف ٨ UTF-8.

ويتيح البرنامج إمكانية قيام الباحثين والمستفيدين بتحميل ومعالجة ملفٍ يحتوي على أكثر من ٥٠ مليون كلمةٍ، من خلال العمل في بيئة جافا JAVA. ويوضح الشكل رقم (٢٩) واجهات الاستخدام لبرنامج أدوات معالجة المدونات اللغوية العربية.

١- يمكن تحميل البرنامج من خلال الرابط التالي:
https://www.researchgate.net/profile/Abdulmohsen_Al-thubaity/publication/310620750_Ghawwas_V46_An_Arabic_Corpora_Processing_Syste m_ghwas_46_nzam_lmaljt_almdwnat_alrbyt/data/58341b5f08aef19cb81da1ff/Ghawwas-V46.jar



الشكل رقم (٢٩) واجهات الاستخدام لبرنامج أدوات معالجة المدونات اللغوية العربية

ويحتوي البرنامج على ثلاث واجهات استخدام:

١. يُمكن من خلال واجهة الاستخدام الأولى أن يقوم الباحث بتحميل ملفات النصوص، سواءٌ الخاصة بالمدونة اللغوية الرئيسة، أو المدونة اللغوية المرجعية.

٢. بينما توفر واجهة الاستخدام الثانية مجموعةً من الإمكانات، هي: تحليل المتتابعات اللفظية n-grams، ومحرك بحث المدونة اللغوية، وحذف أو إبقاء علامات التشكيل، أو تعديل بعض التمثيلات، وتحديد الملفات التي يرغب المستفيدون في البحث فيها، ورفع قوائم الاستثناء أو قوائم الاعتبار.

٣. فيما تعرض واجهة الاستخدام الثالثة مجموعةً من الإحصاءات والتحليلات. حيث يمكن حساب قيمة مربع كاي للدلالة الإحصائية Chi-square، وطريقة احتمالات سجل الأداء log-likelihood، ومعامل ارتباط الغرابة Weirdness Coefficient، ومعامل المعلومات المتبادلة Mutual Information، ومعامل ارتباط دايس Dice Coefficient.

إذ يتم تقييم ملفات النصوص بحيث يتم تمثيل الدلالة الإحصائية لتوزيع كلمةٍ معينةٍ. وتُستخدم طريقة احتمالات سجل الأداء في المقارنات التي تُقدر قيم الاحتمالات، ومن ثم يمكن تقدير قيم معاملات الارتباط. حيث يُشار إلى ذلك بنتائج عاليةٍ ومنخفضةٍ ضمن اختبارٍ معياري.

فيما يُستخدم معامل ارتباط الغرابة في استخراج الكلمات المفتاحية والمتلازمات اللغوية من النصوص؛ من خلال المضاهاة بين مدونةٍ لغويةٍ رئيسةٍ، ومدونةٍ لغويةٍ مرجعيةٍ. ويُشار إلى ناتج معامل ارتباط الغرابة بأربعة قيم؛ القيمة الأولى للمستوى ٠، والقيمة الثانية أكبر من ١ حينما تكون الكلمات أكثر ترددًا في المدونة اللغوية الرئيسة عنها في نظيرتها المرجعية. والقيمة الثالثة أقل من ١ عندما تكون الكلمات أكثر ترددًا في المدونة اللغوية المرجعية عنها في نظيرتها الرئيسة. بينما القيمة الرابعة «إلى ما لا نهاية» حينما ترد الكلمات في المدونة اللغوية الرئيسة فقط.

ويفيد معامل المعلومات المترابطة في التحقق من قوة الارتباط بين المتلازمات اللغوية. فكلما ازدادت القيمة، ازدادت قوة الارتباط بين المتلازمات اللغوية. وتشير عادةً القيمة الأقل من ٣ إلى انعدام الارتباط بينها.

بينما يدل معامل ارتباط دايس على قوة الارتباط أو ضعفه بين الكلمات والوثائق. وتتراوح قيم هذا المعامل بين ٠ و ١٤. وتظهر هذه القيم في حالة وجود ارتباطٍ بين المتلازمات اللغوية. وكلما اقتربت القيمة من ١٤، ازدادت قوة الارتباط. ويوضح الشكل رقم (٣٠) طريقة عرض التحليلات الإحصائية في برنامج غَوَّاص.

في	WF1	WF2	WRF1	WRF2	File1	File2	χ2 1	χ2 2	MI_1	MI_2	z- 1
	13	22	3.74	4.24	S1	S2	0.13	0.0	-0.11	0.06	-0.27
	z- 2	t- 1	t- 2	Dice1	Dice2	LogD1	LogD2	LL1	LL2	WC1	WC2
	0.22	-0.28	0.2218	0.06	0.07	10.12	10.34	0.13	0.13	0.88	1.0
[illegible]	WF1	WF2	WRF1	WRF2	File1	File2	χ2 1	χ2 2	MI_1	MI_2	z- 1
	8	7	2.30	1.35	S1	S2	1.11	0.0	0.41	-0.35	0.80
	z- 2	t- 1	t- 2	Dice1	Dice2	LogD1	LogD2	LL1	LL2	WC1	WC2
	1.0	-0.66	0.701	-0.74	0.04	0.02	9.50	8.74	1.08	1.08	1.70
[illegible]	WF1	WF2	WRF1	WRF2	File1	File2	χ2 1	χ2 2	MI_1	MI_2	z- 1
	8	0	2.30	0.0	S1		12.05	NAN	1.31	0.0	2.67
	z- 2	t- 1	t- 2	Dice1	Dice2	LogD1	LogD2	LL1	LL2	WC1	WC2
	NAN	-2.18	1.69	Infinity	0.04	0.0	9.52	Infinity	14.7	14.72	Infinity
[illegible]	WF1	WF2	WRF1	WRF2	File1	File2	χ2 1	χ2 2	MI_1	MI_2	z- 1
	5	0	1.44	0.0	S1		7.50	NAN	1.31	0.0	2.11
	z- 2	t- 1	t- 2	Dice1	Dice2	LogD1	LogD2	LL1	LL2	WC1	WC2
	NAN	-1.73	1.33	Infinity	0.02	0.0	8.86	Infinity	9.17	9.17	Infinity
[illegible]	WF1	WF2	WRF1	WRF2	File1	File2	χ2 1	χ2 2	MI_1	MI_2	z- 1
	5	0	1.44	0.0	S1		7.50	NAN	1.31	0.0	2.11
	z- 2	t- 1	t- 2	Dice1	Dice2	LogD1	LogD2	LL1	LL2	WC1	WC2
	NAN	-1.73	1.33	Infinity	0.02	0.0	8.86	Infinity	9.17	9.17	Infinity

الشكل رقم (٣٠) طريقة عرض التحليلات الإحصائية في برنامج غَوَّاص

يُذكر أن هناك برامج أخرى يمكنها التعامل مع المدونات اللغوية العربية، مثل: آكونكورد aConCorde (Roberts, 2014)، وآنتكونك AntConc (Anthony, 2005)، واستقصاء المدونات إنتيلي تيكست IntelliText Corpus Queries (Sharoff, 2011).

الخلاصة

قدم هذا الفصل نماذج فعليةً للمدونات اللغوية المعتمِدة على الشبكة العنكبوتية في إتاحتها وتقديم خدماتها. مع إبراز العربية منها قدر الإمكان. ومن بين هذه النماذج: المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية، والمدونة اللغوية التاريخية للجامعة الأردنية، والمدونة اللغوية العربية الدولية لمكتبة الإسكندرية، ومدونة عربي كوربَص، والمدونة اللغوية لمتعلمي اللغة العربية، والمدونة العربية القرآنية، ومدونة قرآني، واستقصاء المدونات اللغوية العربية، وسكتش إنجين، ومدونة كوكا، والمدونة اللغوية لكتب جوجل، ومدونة ناو. كما قدم الفصل أيضًا أشهر البرمجيات التي تعالج وتحلل المدونات اللغوية العربية، ومنها: برنامج أدوات وورد سميث، وبرنامج أدوات معالجة المدونات اللغوية العربية.

الفصل الخامس

دراسات استخدام المدونات اللغوية

التمهيد
دراسات إنشاء وإتاحة المدونات اللغوية العربية
دراسات الإفادة من المدونات اللغوية العربية في علم اللغة التطبيقي
في النحو والدلالة
في علم اللغة الاجتماعي
في صناعة المعاجم
في الترجمة
في دراسة التوجهات الفكرية (الأيديولوجيا)
دراسات استخدام المدونات اللغوية العربية في استرجاع المعلومات
دراسات الإفادة من المدونات اللغوية في صناعة المكانز
دراسات استخدام المدونات اللغوية في المكتبات
الخلاصة

التمهيد

استعرضنا في الفصول السابقة المدونات اللغوية من حيث ماهيتها وأهميتها، ومميزاتها، ومواصفاتها، وما يمكن أن توفره، واستثمارها في مجالات المعرفة المختلفة، وطرائق وأساليب معالجة البيانات القائمة عليها، ونماذجها الفعلية وبرمجيات معالجتها. وبعد كل ذلك وجد المؤلف أن يختتم بالدراسات العلمية، والمقالات البحثية التي وظفت بالفعل تلك الأداة في منهجياتها؛ سواءٌ بالتصميم والإنشاء والإتاحة، أو بالتطبيق العملي.

وفيما يلي عرضٌ لنماذج منتقاةٍ من الدراسات، موزعةً على خمس فئاتٍ رئيسةٍ. وداخل كل فئةٍ رُتبت الدراسات ترتيبًا زمنيًا من الأقدم للأحدث على النحو التالي:

دراسات إنشاء وإتاحة المدونات اللغوية العربية

نود قبل الخوض في استعراض هذه الدراسات أن نشير بدايةً إلى إمكانية التحقق من مراحل إنشاء المدونات اللغوية بوجهٍ عامٍ ضمن تناولها تفصيلًا في الفصل الثاني.

ومن أوائل الدراسات التي تناولت إنشاء وإتاحة المدونات اللغوية العربية دراسةُ عبد الباسط قويدر Goweder و دي روك De Roeck في عام ٢٠٠١ التي وصفت كيفية بناء مدونةٍ لغويةٍ عربيةٍ قوامها ١٨,٥ مليون كلمةٍ من نصوص الأخبار الواردة في جريدة الحياة المكتوبة بلغة تهيئة النصوص الفائقة HTML، والتي تمثل ٤٢٥٩١ مقالةً تغطي ٧ فئاتٍ موضوعيةٍ (عامةٌ، وأخبارٌ، واقتصادٌ، ورياضيةٌ، وحاسبٌ آليٌ، وإنترنت، وعلومٌ وتكنولوجيا، وسياراتٌ وإدارة أعمالٍ). حيث وصفت الدراسة الإطار العام لخصائص وسمات البيانات وكيفية تمثيلها في المدونات اللغوية. وقد حاولت الدراسة التحقق من مدى الاختلاف بين إنشاء وإتاحة المدونات اللغوية العربية والمدونات اللغوية الإنجليزية. وفي هذا السياق استخدم الباحثان قانون زيف Zipf's Law في التوزيعات التكرارية لكل فئةٍ موضوعيةٍ، وفي التوزيعات التكرارية لهذه الفئات الموضوعية مجتمعةً. وقد توصلت الدراسة إلى أنه ليس هناك دليلٌ على افتقاد المدونة اللغوية للتوازن، سواءٌ من حيث التوزيع التكراري أو من حيث مدى الخصوصية (الصفات المميزة للمحتوى). إلا أن النصوص العربية جاءت أكثر تشتتًا من النصوص الإنجليزية، الأمر الذي قد يؤثر على نجاح الطرق المعيارية (مثل تحليل المتتابعات اللفظية n-grams) المطبقة على البيانات العربية.

وفي عام ٢٠٠٥ تعاون أحمد عبدالعالي مع كاوي وسليمان (Abdelali; Cowie & Soliman) في تقديم نموذجٍ لمدونةٍ لغويةٍ عربيةٍ، أو بالأحرى مجموعةٍ فرعيةٍ من المدونات اللغوية العربية، التي من شأنها تيسير سُبل دراسة العربية الفصحى والمقارنة بين اللغة والأسلوب المستخدم في مختلف أقطار الوطن العربي. ومن ثم إمكانية التحقق من التنوع المعجمي والدلالي للمفردات في تلك الأقطار. ومن أجل تحقيق هذا الهدف

قام الباحثون بحصر وجمع نصوص مجموعةٍ من الصحف العربية التي تتيح محتواها على الشبكة العنكبوتية، مستبعدين تلك الصحف التي تتيح مقالاتها في شكل ملفات بي دي إف pdf؛ حيث لا يمكن معالجة النصوص العربية المتاحة بهذا الشكل. ولذلك استبعد الباحثون الكثير من الصحف الشهيرة من العينة المستهدفة، واستبدلوها بصحفٍ أخرى أقل ذيوعًا. وبناءً على ذلك فقد استقر الباحثون على هذه الصحف: الأهرام المصرية، والرأي العام الكويتية، والوطن العمانية، ووكالة الأنباء الجزائرية، وسفير اللبنانية، والجزيرة السعودية، والمغرب اليوم المغربية، وبتراء الأردينة، وراية القطرية، وتشرين السورية، ووكالة الأنباء العراقية.

وقد استخدم الباحثون أحد البرامج الآلية للزحف في الشبكة العنكبوتية يدعى URSA لتجميع محتوى مواقع هذه الصحف. ثم قاموا بتطبيق قانون زبف Zipf's law وصيغة ماندلبورت Mandelbort formula للخروج ببعض المؤشرات الإحصائية حول المدونة اللغوية التي تم جمعها. وقد شملت هذه المؤشرات الإحصائية عددَ ونسبةَ الكلمات مقابل عدد الملفات. واختتم الباحثون دراستهم بالتأكيد على أهمية الاستفادة من محتوى الشبكة العنكبوتية كونه مصدرًا مثاليًا في بناء وتنقيب المدونات اللغوية العربية بغرض دراسة وتحليل الظواهر اللغوية المختلفة التي تصب في صالح دراسات استرجاع المعلومات والترجمة الآلية ومعالجة البيانات اعتمادًا على لغةٍ ونصوصٍ واقعيةٍ.

ثم جاءت لطيفة السليطي وإيريك آتويل ليقدما نموذجًا آخر لمدونةٍ لغويةٍ أسموها «المدونة اللغوية للعربية المعاصرة Corpus of Contemporary Arabic»، وذلك ضمن تقريرٍ مقدمٍ لجامعة ليدز بإنجلترا في عام ٢٠٠٣، ثم أعادا نشره في عام ٢٠٠٦ (Al-Sulaiti & Atwell). حيث أبرز الباحثان في معرِض حديثهما عن الموضوع الدوافع وراء القيام بهذه المهمة، ومدى احتياج اللغة العربية لمثل هذا المشروع، على غرار ما يتم في اللغات الأوربية، وعلى النحو الذي يكفل الوصول الحر للنصوص الواقعية authentic.

وكانت القاعدة الأساس التي بنى عليها الباحثان هذه المدونة اللغوية هي أنها ينبغي ألاَّ تحوي نصوص العربية القياسية فقط، بل أيضًا نصوص العربية المعاصرة المستخدَمة في أمور الحياة اليومية. وقد نظر الباحثان إلى العربية المعاصرة على أنها اللغة المستخدَمة

في مختلف الأقطار العربية، سواءٌ أكانت مكتوبةً أو منطوقةً، منذ بداية تسعينيات القرن الماضي وحتى تاريخ إنشاء المدونة، إضافةً إلى الأنواع اللغوية varieties الإقليمية المعاصرة. وبذلك، وفق رؤيتها، تكون هذه المدونة اللغوية مصدرًا غنيًا للباحثين ولدارسي اللغة العربية تُمكِّنهم من التحقق من العربية القياسية الحديثة، وما يصاحبها من مفرداتٍ جديدةٍ، وأنواعها الإقليمية المختلفة.

ومن أجل تحقيق هذا الهدف قام الباحثان بتوزيع استبيانٍ على المهتمين باللغة العربية من مدرسين ومهندسي لغةٍ؛ بغرض استطلاع الآراء حول الأنواع الأدبية genres التي يمكن أن يتم تضمينها في المدونة اللغوية المزمعة، وكذلك التطبيقات اللغوية التي يمكن أن تفيد فيها هذه المدونة اللغوية. وقد استقر الباحثان وفق ذلك على جمع نصوصٍ عربيةٍ مكتوبةٍ ومنطوقةٍ في موضوعاتٍ مختلفةٍ مستقاةٍ من مجلاتٍ، ومواقع إلكترونيةٍ،وصحفٍ، وقنواتٍ إذاعيةٍ، ورسائل إلكترونيةٍ (emails). وبلغ الحجم النهائي للمدونة اللغوية ٨٤٣٠٠٠ هيكل كلمةٍ.

وفي عام ٢٠٠٧ قدم ياسين بن عجيبة Yassine Benajiba و باولو روسو Paolo Rosso لأول مرةٍ مقياسًا إحصائيًا لتقييم المدونات اللغوية العربية بالاعتماد على قانون زيف. وقد تم بناء هذا المقياس بالاعتماد كليةً على بياناتٍ إحصائيةٍ. وقد نوه الباحثان عن أنه ليس بالضرورة أن ينجح تطبيق هذا المقياس على المدونات اللغوية للغات الأخرى. واختار الباحثان لتطبيق هذا المقياس أربع مدوناتٍ لغويةٍ في موضوعاتٍ مختلفةٍ لتحديد خصائص كل مدونةٍ لغويةٍ منها. تكونت المدونة اللغوية الأولى من ٦٦٠٠٠ هيكل كلمةٍ (أكثر من ٣٦٠ كيلو بايت) من شعر أبي الطيب المتنبي. وتكونت المدونة اللغوية الثانية من ٥٠٠٠٠ هيكل كلمةٍ (حوالي ٢٦٠ كيلو بايت) وردت في ١١١ مقالةً خبريةً. فيما تكونت المدونة اللغوية الثالثة من ٥٥٠٠٠ هيكل كلمةٍ (حوالي ١٢٦ كيلو بايت) وردت في أحد الكتب الدراسية. بينما تكونت المدونة اللغوية الرابعة من حوالي ٦٥٠٠٠ هيكل كلمةٍ (أكثر من ٤٦٠ كيلو بايت) مأخوذةٍ من كتب الإمام ابن القيم الجوزية.

وقد اختبر المقياس المستخدم في هذه الدراسة ثلاثة عوامل رئيسةٍ في كل مدونةٍ لغويةٍ، هي: مدى التعقيد، ومدى التنوع، ومدى صحة التوزيع التكراري لكلمات المدونة اللغوية. وتوصلت النتائج الأولية إلى وجود علاقة ارتباطٍ بين أسلوب الكتابة

وطبيعة النصوص. وقد سجل مستوى التعقيد قِيَمًا مرتفعةً في المدونات اللغوية التي تركز بشكلٍ أكبر على المحتوى أكثر منه في أسلوب الكتابة، كما أن التنوع في استخدام المفردات اللغوية كان أقل في المدونة العلمية.

وفي عام ٢٠١٠ قدم أشرف عبد الرؤوف وزملاؤه (AbdelRaouf & et al.) ورقةً بحثيةً قدموا فيها وصفًا لبناء مدونةٍ لغويةٍ ذات أنواعٍ أدبيةٍ متعددةٍ قوامها ٦ مليون هيكل كلمةٍ مُجمَّعةٍ من مصادر مختلفة:

١. المواقع الإلكترونية ذات الموضوعات المتعددة.
٢. المواقع الإلكترونية للصحف والجرائد.
٣. غرف الدردشة الإلكترونية.
٤. المعاجم العربية العربية.
٥. الكتب العربية القديمة.
٦. الإنتاح الفكري الأكاديمي.
٧. القرآن الكريم.

ووصف الباحثون النصوصَ التي حصروها بأنها: «عربيةٌ قديمةٌ»، ونصوصٌ دينيةٌ، و «لغةٌ تقليديةٌ»، و «لغةٌ حديثةٌ»، و تخصصاتٌ متنوعةٌ، ومصادر «حديثةٌ جدًّا» مستقاةٌ من غرف الدردشة الإلكترونية. وتتميز هذه المدونة اللغوية بأنها مزودةٌ بالصور الأصلية للوثائق العربية التي تمت رقمنتها؛ بغرض الإفادة منها في عمليات السحب الضوئي لها.

كما نشرت سلوى حمادة في عام ٢٠١١ مقالةً عن المدونات اللغوية العربية، تناولت فيها أهميتها في حل المشكلات اللغوية، وصناعة المعاجم، وتعليم اللغات. كما أشارت الباحثة إلى المشكلات التي تعوق إنشاء المدونات اللغوية الخاصة باللغة العربية، وكيفية مواجهتها، وكيفية جمع النصوص الخام التي تشكل البنية الأساس لعمل المدونات اللغوية. ثم ساقت الباحثة مجموعةً من النماذج الفعلية لمدوناتٍ لغويةٍ وبرمجيات تحليلها. وتطرقت المقالة في جزئها الثاني إلى تحديد أهم النقاط التي يجب مراعاتها عند إنشاء المدونات اللغوية، وخطوات عملها، وكيفية وضع التحشية بها، وكيفية تشفيرها.

وفي عام ٢٠١٣ نشر محمد عبد المجيد منصور بحثًا في المجلة الدولية للإنسانيات والعلوم الاجتماعية أكد فيه على الدور الكبير الذي تؤديه المدونات اللغوية في معالجة

البيانات والتحليل والإحصاء اللغوي. كما أشار إلى أنه على الرغم من الأهمية المتزايدة للمدونات اللغوية، إلا أننا نفتقد لمثل هذا النوع من مصادر المعلومات في وطننا العربي. وفي إطار محاولته لوضع حلولٍ عمليةٍ اقترح الباحث أنموذجًا لمدونةٍ لغويةٍ عربيةٍ أَطلق عليها اسم «المدونة القومية العربية» Arabic National Corpus (ANC)، وذلك وفق أربع مراحل من التصميم: المرحلة الأولى «التخطيط للمدونة اللغوية»، والمرحلة الثانية «تجميع البيانات» (النصوص المكتوبة والنصوص المنطوقة)، والمرحلة الثالثة «حوسبة المدونة اللغوية»، والمرحلة الرابعة «تحليل المدونة اللغوية». وقد أكد الباحث أيضًا أن مثل هذا المشروع يفوق مستوى الأفراد؛ فهو يحتاج إلى تضافر جهود مؤسساتٍ عدةٍ، إضافةً إلى دعمٍ ماديٍ وحكوميٍ كبيرين، وتعاون إقليميٍ بين بلدان الوطن العربي.

كما قدم الميمان ولي (Almeman & Lee, 2013) نموذجًا لمدونةٍ لغويةٍ عربيةٍ متعددة اللهجات باستخدام الشبكة العنكبوتية مصدرًا أساسًا للنصوص. وقد تضمنت منهجيتهما خمس خطواتٍ رئيسةٍ لتجميع وبناء المدونة اللغوية.

انطوت المرحلة الأولى على تجميع الكلمات والعبارات متعددة اللهجات؛ أي التي تُستخدم في بلدانٍ عربيةٍ مختلفةٍ. وهذه اللهجات هي: الخليجية، والمصرية، والشمال أفريقية، والشامية. وفي هذا السياق لجأ الباحثان إلى الشبكة العنكبوتية لاستخراج كلمات اللهجات. وفي النهاية بلغ إجمالي هذه الكلمات ١٥٠٠ كلمةٍ تم تصنيفها في قوائم وفقًا لكل لهجةٍ.

فيما انصبت الخطوة الثانية على التمييز بين الكلمات التي تم جمعها، كلٌ وفق لهجتها. وهنا قام الباحثان بالتحقق من لهجة كل كلمةٍ بالاستعانة بستة محكمين. إذ اضطلع كل محكم بالتحقق من كلمات كل قائمةٍ، فيما عدا القائمة التي تخص لهجته الأم. وبعد تنقيح الكلمات تم الاستقرار على ١٠٤٣ كلمةً موزعةً على اللهجات الأربع.

وقبل أن تتم خطوة التحميل قام الباحثان بحساب متوسط عدد هياكل الكلمات tokens التي سيتم إنتاجها لكل رابط أو صفحة إنترنت. وهذا ما تضمنته الخطوة الثالثة. وبعد تقدير عدد الصفحات المطلوب تحميلها لكل لهجةٍ تم تنفيذ الخطوة الرابعة: خطوة التحميل. حيث استعان الباحثان هنا بواجهة بنج إيه بي آي Bing API لتحميل الصفحات، ثم حفظها في شكل ملفات إتش تي إم إل html.

وقام الباحثان في الخطوة الخامسة والأخيرة بتهذيب وتوحيد النتائج التي تم جمعها من مصادرها المختلفة: المنتديات الإلكترونية، والمدونات، وتعليقات المستفيدين...

إلخ. واشتملت هذه الخطوة على التخلص من الرموز والتيجان، والمسافات الزائدة... غير المرغوب فيها جميعًا. كما اشتلمت أيضًا على التخلص من الكلمات والعبارات المكرَرَة في الصفحات التي تم تجميع محتواها، وخاصةً الكلمات شائعة الاستخدام، مثل: الصفحة الرئيسة، وعضو، والتسجيل... إلخ.

وبلغت الحصيلة النهائية للمدونة اللغوية التي تم تجميعها ٥٠ مليون هيكل كلمةٍ، منها ٢ مليون كلمةٍ فريدةٍ type. وفي النهاية تم التحقق من جودة هذه المدونة اللغوية من خلال مقارنتها بمدونتين لغويتين أخريين، لاستكشاف العيوب، وإبراز المميزات. وفي الختام أوصى الباحثان بإمكانية الاعتماد على المدونة اللغوية المقدَمة في معالجة اللغة الطبيعية لنصوص اللهجات العربية.

فيما قام إبراهيم أبو الخير (Abu El-Khair, 2016) ببناء مدونةٍ لغويةٍ للأخبار المنشورة بمواقع ١٠ صحفٍ عربيةٍ من ٨ بلدانٍ عربيةٍ، قوامها ٣٣٠٣٧٢٣ هيكل كلمةٍ. ولتحقيق ذلك استخدم الباحث برنامجين لاستخراج النصوص من هذه المواقع؛ هما: برنامج (5)MetaProducts Offline Explorer Pro، و Visual Web Ripper. كما أوضح الباحث أسباب اختياره لهذين البرنامجين دون غيرهما؛ وهي أنهما أسرع في الاستخدام، فضلًا عن أنهما يتيحان إمكانية استخراج النصوص فقط دون الكيانات الأخرى غير الضرورية، كالصور، وملفات الفيديو، وملفات جافا سكريبت JavaScript files، وملفات سي إس إس CSS files.

وإضافة إلى ذلك، فقد وضع الباحث لنفسه مجموعةً من المعايير لاختيار مواقع الأخبار العربية العشرة هذه، منها: ألَّا يكون هناك تكرارٌ لأي مجهودٍ بُذل في إنشاء مدوناتٍ لغويةٍ من قبل، وأن يكون الموقع متاحًا باستمرار، ويسمح بالزحف[١] crawling في محتواه، ومن ثم يمكن استخراج نصوصه بسهولةٍ ويُسرٍ، وأن تكون عينة المواقع ممثلةً لدولٍ عربيةٍ مختلفةٍ، وأن تكون النصوص قابلةً للتحرير.

وقد لجأ أبو الخير إلى ترميز marking-up مدونته بإضافة حقول الميتاداتا مستخدمًا لغة الترميز المُعَمَمَة القياسية SGML ولغة الترميز القابلة للتمديد XML لترميز النصوص. كما أنه قام بتشفير النصوص وفقًا لنظام يو تي إف ٨ UTF-8.

١- هي عملية تعيين واستدعاء واختزان وثائق الشبكة العنكبوتية بواسطة برامج مخصصة لذلك.

دراسات الإفادة من المدونات اللغوية العربية في علم اللغة التطبيقي

كثيرةٌ هي الدراسات القائمة على المدونات اللغوية العربية في علم اللغة التطبيقي. ومن غير المبالغ فيه إذا قلنا إن علم اللغة التطبيقي يُعد الأوفر حظًا بين علوم ومجالات المعرفة الأخرى المستفيدة من المدونات اللغوية.

ومن الصعوبة بمكانٍ أن نشير إلى كافة الدراسات التي استثمرت المدونات اللغوية العربية في هذا الشأن. فهذا الطرح جديرٌ بأن يُفرد له كتابٌ لحصر الإنتاج الفكري حول الموضوع بحيث يكون مرجعًا لكل المهتمين. غير أننا سنحاول هنا أن نسرد بعضًا من هذه الدراسات محاولين تصنيفها تبعًا لفروع علم اللغة المختلفة.

في النحو والدلالة

نود قبل الخوض في استعراض هذه الدراسات أن نشير بدايةً إلى إمكانية التحقق من الإفادة من المدونات اللغوية في تدريس وتعلم اللغات بوجهٍ عامٍ ضمن تناولها تفصيلًا في الفصل الثاني.

وفي هذا السياق، استعانت شيماء عيسى (Essa, 2013) خلال رسالتها للماجستير بمدونة عربي كوربص أداةً رئيسةً للبحث والدراسة، وتحديدًا، المدونة اللغوية الفرعية لأعمدة الشروق البالغ عدد كلماتها ٢٠٦٧١٣٧ هيكل كلمةٍ، لدراسة ثلاثٍ من أدوات الربط العربية؛ وهي: إنما، وبينما، وبَل. وذلك بهدف التحقق من سلوكها اللغوي كونها أدوات استدراكٍ adversative. واكتفت الباحثة بعينةٍ عشوائيةٍ لـ «بل» و «إنما» تغطي ٥٠٪ من الأمثلة السياقية المذكورة في هذه المدونة اللغوية. فيما أخذت الباحثة كافة الأمثلة المذكورة لـ «بينما» لتحليلها؛ نظرًا لقلة عدد أمثلتها المذكورة بالمدونة اللغوية التي تحمل المعنى الاستدراكي.

وانصب التحليل في هذه الدراسة على أدوات الربط الثلاث هذه في ضوء التركيب المعجمي النحوي lexico-grammatical pattern، والسلوك التلازمي collocational behavior، والعروض الدلالي semantic prosody. وأوضحت النتائج تشارك «إنما» و «بل» في بعض التراكيب المعجمية النحوية، والسلوك التلازمي. فيما تفردت «بينما» بخصائصها. ولم يكن هناك لـ «إنما» و «بينما» عروضٌ دلاليٌ واضحٌ، سواءٌ في المعنى الإيجابي أو في المعنى السلبي لهما. وذلك على عكس أداة الربط «بل» التي

كان لأحد تراكيبها المعجمية النحوية عروضٌ دلاليٌ واضحٌ. وأظهرت النتائج كذلك أن أدوات الربط الثلاث لها السلوك التلازمي نفسه..

وأشارت الدراسة أيضًا إلى أنه على الرغم من أن بعض السلوكيات اللغوية لأدوات الربط الثلاث تعزو إلى أوجه التشابه أو الاختلاف في معانيها، فإن بعض السلوكيات اللغوية الأخرى لا تُفهم إلَّا من خلال الكشف عن درجة دلالة هذه الأدوات في زمن المستقبل. كما أكدت الدراسة على الدور الذي يمكن أن تُسهم به نتائج هذه الدراسة في تدريس اللغة العربية؛ سواءٌ في إعداد الدروس التعليمية، أو في تحديث وتطوير المواد والكتب الدراسية حول أدوات الربط الثلاث هذه.

فيما قدم أحمد إسماعيل (Ismail, 2015) في رسالته للماجستير منهجيةً اعتمد فيها على مدونةٍ لغويةٍ قوامها ٨٦٥٩٢ هيكل كلمةٍ لَمْتن سبعة أفلامٍ سينمائيةٍ مصريةٍ؛ هي: عمارة يعقوبيان، وبحب السيما، وأرض الخوف، والكيت كات، والبداية، والكرنك، وفي بيتنا رجل. وذلك بهدف دراسة ثلاثٍ من أدوات الخطاب العامِّي في الدَارِجة المصرية؛ هي: بَقَىَ، وطَيِّبْ، وطَبْ. واعتمد الباحث في تحليله للمدونة اللغوية على برنامج أدوات وورد سميث.

واقتصر الباحث في دراسته على تحليل النواحي الدلالية والنحوية والتلازمية لهذه الأدوات الثلاث. وقد أظهر التحليل أن هذه الأدوات الثلاث تؤدي وظائف لغويةً عدةً، فضلًا عن أنه يمكنها أن تعمل (بشكلٍ متزامنٍ أحيانًا) على مستوى الخطاب والمحادثات الشخصية. كما أنها لا تملأ فراغات الكلام فحسب، بل أيضًا تؤدي دورًا حيويًا في بناء وتماسك العلاقات النصية والاجتماعية للكلام.

وإضافةً إلى ذلك، فإن سطور كشاف نصوص كلمة «بَقَىَ» أوضحت أنها تفيد في إضفاء السلاسة والتماسك على لغة الخطاب من خلال دورها في تيسير وتوضيح وإنهاء الكلام، وإشارتها لتوجهات وشعور المتكلم. وبينما تُستخدم «بَقَىَ» في بدء الحديث، فإن كلمة «طَيِّبْ» وكلمة «طَبْ» تتخذان مواقع مختلفةً داخل الجمل، وذلك على حسب الوظائف النحوية التي تؤديها هاتان الأداتان الخطابيتان.

كما تقدم سلطان المجيول (٢٠١٦) بورقةٍ بحثيةٍ للمؤتمر الدولي الثاني في الاتجاهات الحديثة في تعليم العربية لغةً ثانيةً، اقترح من خلالها تحليلًا تطبيقيًا لنظرياتٍ معجميةٍ في إطار التحليل الآلي المدوَّني في لسانيات المدونات اللغوية Corpus Linguistics،

ولأغراض التحليل المعجمي المفتوح للغة، وفقًا لسياقها البحثي المتنوع والخاص لغرضٍ معجميٍ معيّنٍ. وركزت الورقة على مناهج نظريةٍ تطبيقيةٍ تعُرف باسم: التهيئة المعجمية lexical priming والتلازم اللغوي collocation ودور «التجميع» nesting وتعبئة الفجوة المعجمية lexical gap-filling في النحو النمطي pattern grammar. وسعى الباحث في ورقته هذه إلى تطبيق هذه الاتجاهات على معجم العربية بين يديك للناطقين بغير العربية (الفوزان وآخرون ١٤٢٥) من جهةٍ، والمعجم الوسيط للناطقين بالعربية (٢٠٠٤) من جهةٍ ثانيةٍ؛ وذلك بهدف الكشف عن مدى حضور التعبئة المعجمية وتهيئتها في متون هذه المعاجم، وعن فوائد تلك المناهج في تطوير المعجم العربي الخاص بتعليم العربية لغير الناطقين بها. وأورد الباحث سبب إدخال المعجم الوسيط إلى كونه معاصرًا ومتسعًا نوعًا ما، وجامعًا بين الشارد وبين الجِدَة في الاستعمال أولًا، ومتوفرًا -ثانيًا- لتحويله إلى ملف النصوص البسيطة plain text، وملف القيم المفصولة بفاصلة (إكسل) comma separated value csv.

بينما استخدم أيمن الدكروري (Eddakrouri, 2016) المدونة اللغوية أوساك OSAC، التي تحوي نصوص الأخبار العربية المنشورة بموقعي سي إن إن CNN و بي بي سي BBC، والبالغ عدد كلماتها ٤١٠٢١٣٤ هيكل كلمةٍ، وذلك لدراسة وتحليل اثنين من أكثر الظُرُوف العربية استخدامًا؛ وهما: بعد، وقبل.

وقد اعتمد الباحث في تحليله لهذه المدونة اللغوية على برنامج أدوات وورد سميث لاسترجاع كافة السياقات التي ورد بها هذان الظرفان. وقد انضوى التحليل اللغوي لهذه السياقات على الخصائص الدلالية والوظائف الترابطية، والمميزات النحوية، والاستخدامات الاصطلاحية لهذين الظرفين. ومن ثم تمكن الباحث من تقسيم واقعات تكرار كلمة "بعد" إلى ثلاث فئاتٍ رئيسةٍ تبًعا لمعناها؛ وهي: بعد (في معناها المباشر)، وبعد ذلك/ ثم، والتعبيرات الاصطلاحية لـ "بعد". كما تم تقسيم واقعات تكرار كلمة "قبل" إلى ثلاث فئاتٍ أيضًا، هي: قبل (في معناها المباشر)، ومنذ، والتعبيرات الاصطلاحية لـ "قبل".

وقد أبرزت نتائج الدراسة الدور الكبير الذي يؤديه هذان الظرفان في وحدة، وتماسك، وانتظام، وفهم الأخبار العربية المنشورة على الإنترنت. كما كشفت نتائج الدراسة عن المعاني الواقعية الأخرى لهاتين الكلمتين؛ كتلك التي تم توضيحها عند

تقسيمها إلى فئاتٍ تبعًا لمعانيها ودلالتها. وكان من بين النتائج أيضًا ما أبرزته الدراسة من وجود اختلافٍ في تكرار تردد كل معنىً من معاني «قبل» و «بعد». حيث تُستخدم «بعد» بمعناها المباشر في معظم السياقات التي وردت بها (٨٣.٧٪)، ويأتي بعدها معنى «بعد ذلك/ ثم» (١٤.٥٪)، ثم باقي الاستخدامات الاصطلاحية لها (١.٧٪). كما استُخدِمَت «قبل» بمعناها المباشر في معظم سياقاتها أيضًا (٦٢.٧٪)، ثم بمعنى «منذ» (٢٤.٣٪)، ثم باقي استخداماتها الاصطلاحية الأخرى (٩.٧٪).

في علم اللغة الاجتماعي

نود قبل الخوض في استعراض هذه الدراسات أن نشير بدايةً إلى إمكانية التحقق من الإفادة من المدونات اللغوية في علم اللغة الاجتماعي بوجهٍ عامٍ ضمن تناولها في الفصل الثاني.

وفي هذا الإطار، قدم مارك فان مول (Van Mol, 1998)، صاحب الإسهامات البارزة في هذا الصدد، دراسةً رائدةً في استثمار المدونات اللغوية العربية. كانت هذه الدراسة بالأساس رسالة دكتوراه مكتوبةً باللغة الألمانية، نوقشت في الجامعة الكاثوليكية ببلجيكا. ثم قام فان مول بعدها بحوالي ست سنواتٍ بترجمتها إلى الإنجليزية ونشرها في صورة كتابٍ (Van Mol, 2003). وتمثل الهدف الرئيس من هذه الدراسة في التحقق من التباين variation في استعمال اللغة العربية من دولةٍ لأخرى اعتمادًا على المدونات اللغوية.

حيث اعتمد الباحث على مدونةٍ لغويةٍ موَّسَمَة الكلمات POS tagged corpus لنشرات الأخبار الإذاعية قوامها ٣٢٠٠٠٠ كلمةٍ منسوخةٍ transcripted. واستقر الباحث على لغة الإعلام نوعًا أدبيًا مُمَثِلًا للعربية القياسية الحديثة Modern Standard Arabic (MSA). كما استقر الباحث على ثلاث دولٍ عربيةٍ مختلفةٍ، هي: الجزائر، ومصر، والسعودية للتأكد من مدى التباين اللغوي فيما بينها. ولبلوغ هذا الهدف قام الباحث بوَسْم tagging هذه المدونة اللغوية على مستوى الكلمة اعتمادًا على قواعد النحو العربي.

وقد ركز الباحث في دراسته على وصف الزمن أو التزامن في استخدام الأدوات التكميلية complementary particles بالمرجعية لقواعد نحو العربية المعاصرة. وقد أوضحت نتائج تحليل المدونة اللغوية أن هناك اتساقًا كبيرًا في استخدام الأدوات

التكميلية في هذه الدول العربية الثلاث على المستوى النحوي. غير إن هناك اختلافًا كبيرًا في الوظائف التي حددتها كتب النحو التقليدية للأدوات التكميلية موضوع الدراسة. حيث إن استخدام «س» و «سوف» للاستقبال لم يميز بين المستقبل القريب والمستقبل البعيد، كما أوضحت كتب النحو التقليدية. وأظهرت الدراسة كذلك اختلافاتٍ جغرافيةً في استعمال عبارتي «في الوقت نفسه» و «في نفس الوقت» اللتين وردتا بالعدد نفسه من التكرار. إلّا أنه كان هناك اختلافٌ في استعمالهما في الدول المختلفة. فعبارة «في الوقت نفسه» وردت بكثرةٍ في النصوص المصرية، بينما لم ترد في النصوص الجزائرية. أما في النصوص السعودية فقد وردت العبارتان بتردداتٍ متساويةٍ.

يُذكر أن فان مول طَوَّر المدونة اللغوية التي بناها في الدراسة السابقة، واستثمرها في مجموعةٍ متعاقبةٍ من الدراسات الأخرى التي أفادت علم اللغة التطبيقي بوجهٍ عام.

وقد لحق باحثٌ آخر بقطار الرواد في دراسات علم اللغة الاجتماعي المعتمِدة على المدونات اللغوية، وهو ديلوورث باركينسون Dilworth Parkinson الباحث بجامعة بريجهام يانج الأمريكية Brigham Young University. ذلك حينما قدم دراسةً للتحقق من التباين اللغوي في استخدام الأدوات الدالة على المستقبل future particles بين الدول العربية (Parkinson, 2003). إذ اعتمد باركينسون على مدونةٍ لغويةٍ قوامها ٤٥ مليون كلمةٍ تم تجميعها من صحفٍ عربيةٍ تمثل دولًا مختلفةً، وهي: الأهرام المصرية، والحياة السعودية اللبنانية، والتجديد المغربية، والوطن الكويتية.

وقد أوضحت نتائج تحليل الأدوات الدالة على المستقبل في هذه الدراسة، وهي «س» و«سوف»، أن هناك تباينًا كبيرًا من دولةٍ عربيةٍ لأخرى، ومن نوعٍ أدبيٍ لآخر، ومن فعلٍ لآخر.

كما اعتمد أحمد عبد العالي (Abdeali, 2004) على مدونةٍ لغويةٍ لعشر صحفٍ قوميةٍ، تمثل دولًا عربيةً مختلفةً، قوامها ٢٢٢٨٠١٦ هيكل كلمةٍ؛ بهدف التحقق من توطين اللغة localization من خلال مقارنة اللغة المستخدمة في مواطن مختلفةٍ من العالم العربي، في ضوء الخصائص المعجمية للعربية القياسية الحديثة (MSA) Modern Standard Arabic.

وأوضحت النتائج أن العربية القياسية الحديثة تبدو متسقةً جدًّا في كافة الدول العربية التي شملتها الدراسة. غير إن هناك بعض الفروق أو الاختلافات المهمة التي تم إبرازها، والتي يمكن الاعتداد بها في المعالجة الآلية للغة. وقد قسَّم الباحث هذه

الاختلافات على النحو التالي:

١. اختلافاتٌ في رسم الحروف: فالطبيعة الاشتقاقية العالية للغة العربية أدت إلى اختلافاتٍ ملحوظةٍ في رسم الكثير من الحروف بين البلدان العربية. ومن الأمور الأخرى التي أدت بشكلٍ أو بآخر إلى إحداث هذه الاختلافات المراحل المتعاقبة حول تطوير قواعد النحو العربي، والتلاوات المتعددة لقراءة القرآن الكريم. وتتضح هذه الاختلافات في بعض الكلمات، نحو: إدارة – أدارة، أيام – ايام، الأربعاء – الاربعاء، أحمد – احمد.

٢. اختلافاتٌ في نقحرة الكلمات: فللغات الأجنبية تأثيرٌ على العربية القياسية الحديثة. ويظهر هذا التأثير في طريقة نطق بعض الكلمات اعتمادًا على نطقها باللغة التي استُقيت منها، ولاسيما اللغة الإنجليزية، واللغة الفرنسية. فاختلاف صوت حرف « i » في الإنجليزية عنه في الفرنسية، وكذلك الحال مع أصوات حروف a و e و ch، وغيرها، أدى، ربما، إلى الاختلاف في نقحرة بعض الكلمات، نحو: الإنترنت – الانترنت.

٣. اختلافاتٌ في الاستعمال: فهناك بعض الكلمات التي تختلف في استعمالاتها من منطقةٍ عربيةٍ لأخرى. حتى أن هناك كلماتٍ عربيةً يندر، أو ربما يندثر، استعمالها في بعض أجزاء من الوطن العربي، بينما تظل هذه الكلمات مستخدمةً في أجزاءٍ أخرى. كما في: عنبر – مرقد، حجز - توقيف، آيلة للسقوط – معرضة للسقوط.

٤. اختلافاتٌ في الأسماء: فقد أدت الطبيعة القبلية القوية في بعض المناطق العربية إلى استعمال بعض الأسماء في نطاق بعض القبائل دون خارجها. كما في: فهد، فوزية، موضى، نورة، سطام، آل، الغامدي، العتيبي، الحربي، الشمري، القحطاني، الزهراني.

٥. اختلافاتٌ في نحت واشتقاق الكلمات: تظهر هذه الاختلافات جليًا مع الكلمات الدخيلة التي تختلف باختلاف اللغة المُصَدِرة للكلمة، وعملية التكييف نفسها. كما في: كابل – حبل <<< cable، الخصخصة – الخوصصة <<< Privatization.

واعتمد عبد الرزاق (Abdul Razak, June 2014)، الباحث بجامعة ماليزيا للعلوم الإسلامية، على مدونةٍ لغويةٍ قوامها ٨٧٠٠٠ هيكل كلمةٍ، جمع محتواها من قسم الشؤون الدولية لسبع صحفٍ عربيةٍ مختلفة المَنشَأ؛ وذلك بهدف التحقق من الاختلافات في

استخدام الكلمات العربية من دولةٍ لأخرى. وقد تبنى الباحث نظرية آدم كيلجاريف (Kilgarriff, 2001) التي تذهب إلى أن تحليل مدى الاتساق في استخدام الكلمات في المدونات اللغوية من الممكن أن يكشف عن الاختيارات المعجمية. ولذا ركزت الدراسة على مناقشة أنواعٍ عدةٍ من تباين الأسماء variations of nouns، واستخدام الأفعال.

وفي هذا السبيل، استخدم الباحث في تحليله للمدونة اللغوية برنامج أدوات وورد سميث. وقد أظهرت النتائج أن الصحف العربية لها دورٌ كبيرٌ في إحداث التباين في استخدام الكلمات؛ ولاسيما فيما يخص الهجاء، والكلمات الدخيلة، وتعدي الأفعال، وتكوين العبارات، وبناء صيغ الجمع. وقد أرجع الباحث بعضًا من هذا التباين إلى تأثير اللغات الأجنبية، وخاصةً اللغة الإنجليزية، على استخدام الكلمات في اللغة العربية.

في صناعة المعاجم

نود أيضًا هنا، وقبل الخوض في استعراض هذه الدراسات، أن نشير أولًا إلى إمكانية التحقق من الإفادة من المدونات اللغوية في صناعة المعاجم بوجهٍ عامٍ ضمن تناولها تفصيلًا في الفصل الثاني.

وقد جاءت أول محاولةٍ لاستثمار المدونات اللغوية في صناعة المعاجم العربية على يد مارك فان مول (Van Mol, 2000). ذلك حينما قام هذا الباحث ببناء مدونةٍ لغويةٍ للعربية القياسية الحديثة قوامها ٣٠٠٠٠٠٠ كلمةٍ تغطي اللغة المستخدمة في عددٍ من الدول العربية، من بينها الجزائر، ومصر، والسعودية. حيث قام الباحث بتجميع النصوص المكتوبة في ٥٠ كتابًا دراسيًا. وهكذا مثلت هذه النصوص المكتوبة ما نسبته ٧٥٪ من إجمالي نصوص المدونة اللغوية، والبقية (٢٥٪) نصوصٌ منطوقةٌ.

ثم قام الباحث بترجمة كل كلمةٍ وكل جملةٍ وردت بالمدونة اللغوية، مستعينًا ببعض المعاجم الأخرى، حتى بلغ عدد المداخل العربية ١٧٠٠٠ مدخلٍ، بينما بلغ عدد المداخل الألمانية ٢٠٠٠٠ مدخلٍ.

وبعد هذه المحاولة بعامٍ، قدم فان مول دراسةً أخرى مع زميله هانس بولوسن (Van Mol & Paulussen, 2001) اعتمدا فيها على مدونةٍ لغويةٍ للعربية القياسية الحديثة قوامها ٣٠٠٠٠٠٠ كلمةٍ. ويبدو غالبًا أنها المدونة اللغوية نفسها التي اعتمد عليها فان مول في دراسته السابقة. وأوضح الباحثان أن الهدف من بناء هذه المدونة اللغوية

هو تصميم قاعدة بياناتٍ علاقيةٍ أسماها آرالات AraLat، اختصارًا لـ «العربية مقابل لغات النصوص اللاتينية Arabic vs. Latin script Languages». بحيث يمكن استخدام قاعدة البيانات هذه في توليد نُسخ ورقيةٍ وإلكترونيةٍ للمعاجم العربية ثنائية اللغة. وكان المعجم الناتج عن هذا المشروع معجمٌ عربيٌ-ألمانيٌ، ألمانيٌ-عربيٌ.

واعتمد الباحثان في بناء هذه المدونة اللغوية على نصوصٍ عربيةٍ واقعيةٍ منطوقةٍ. غير أنهما لم يذكرا مصدرها. ومن أجل الخروج بنتائج أكثر دقةٍ، قام الباحثان بِوَسْم tagging كلمات آرالات آليًا. وقد امتد وسم الكلمات ليشمل أيضًا مستوى الجذر؛ بحيث يمكن ترتيب الكلمات في المعجم وفقًا للجذور.

وبعد أن ترجم الباحثان كل الكلمات العربية إلى الألمانية، قاما بتقسيم التراكيب إلى أربعة مستويات من التحليل والعرض. وهكذا يمكن للباحثين أن يستفيدوا من مدونة آرالات عند الشروع في صناعة أي معجمٍ عربيٍ للعربية القياسية الحديثة.

وفي العام نفسه، جاء سالم غزالي وعبد الفتاح برهام (Ghazali & Braham, 2001) بدراسةٍ حاولا فيها التأكيد على الدور الكبير الذي يمكن أن تؤديه المدونات اللغوية في صناعة المعاجم العربية. حيث اعتمد الباحثان على مدونةٍ لغويةٍ قوامها ١٥٠٠٠٠٠ كلمةٍ تم تجميعها من عدة صحفٍ ومجلاتٍ عربيةٍ، في موضوعاتٍ مختلفةٍ، ومن دولٍ مختلفةٍ. وكانت من مصادر هذه المدونة اللغوية أيضًا كتبٌ للمرحلة الثانوية في كافة المواد الدراسية، عدا مادة العلوم، يتم تدريسها في دولة تونس. إضافةً إلى الإنجيل، وأطروحات دكتوراه، وقصصٍ وروايات، مثل «الأيام» لطه حسين، وأعمالٍ أخرى في الأدب العام.

ومن أجل إثبات الفرضية التي حاول الباحثان طرحها، فقد اختارا فعلًا عربيًا شائع الاستخدام، وله معانٍ مختلفةٌ، وهو الفعل «أخذ». حيث سرد الباحثان له المعاني التي وردت في «المعجم الوسيط». ثم حصرا السياقات المختلفة التي ورد بها هذا الفعل في المدونة اللغوية، وصنفا هذه السياقات وفقًا لكل معنىً من معاني «أخذ».

وأوضحت نتائج تحليل معاني «أخذ» أن هناك معانٍ سياقيةً أغفلتها معاجم اللغة العربية. هذه المعاني لا تتضح سوى بالرجوع إلى أكبر قدرٍ ممكنٍ من السياقات، مع تحليل البيئة المعجمية النحوية للكلمة.

كما كان لسامح الأنصاري إسهامٌ في التأكيد على الدور الكبير الذي يمكن أن تؤديه المدونات اللغوية في صناعة المعاجم العربية. ذلك حينما قدم هذا الباحث مقاربةً منهجيةً لبناء معجم آلي للغة العربية اعتمادًا على مدونةٍ لغويةٍ (Al-Ansary, 2005). حيث أوضح الباحث كيفية استثمار المدونات اللغوية في تصميم أداةٍ لمعالجة اللغة الطبيعية، المتمثِلة في المعجم الآلي الذي اقترحه؛ تحقيقًا لأربعة أهدافٍ، هي:

١. تقديم استراتيجية للتحقق من المداخل العربية لمعجمٍ آلي قائمٍ على بياناتٍ واقعيةٍ authentic.
٢. تصميم أداةٍ تُسهم في تحليل مدونةٍ لغويةٍ ضخمةٍ للغة العربية.
٣. رسم الخطوط العريضة لقاعدة بياناتٍ معجميةٍ للغة العربية على غرار قاعدة بيانات سيليكس CELEX للغة الإنجليزية.
٤. إرساء البنية الأساس للتطبيقات العربية لمعالجة اللغة الطبيعية.

ولتحقيق هذه الأهداف فإن الأمر استدعى مجموعةً من المتطلبات، التي أوضحها الباحث في الآتي:

١. مدونةٌ لغويةٌ للعربية القياسية الحديثة.
٢. واجهة استخدام لتحليل النصوص العربية.
٣. نظام إدارة قواعد بياناتٍ لغويةٍ للفحص الآلي للبيانات التي يتم تحليلها.
٤. شكليةٌ لتنفيذ المعجم.

وأرسى الباحث لنفسه مجموعةً من الخطوات أو المراحل للشروع في إنشاء هذا المعجم العربي المقترح، هي:

١. مرحلة تحليل النصوص.
٢. مرحلة الخصائص اللغوية والفئات المعجمية.
٣. مرحلة دمج المدونة اللغوية وقاعدة البيانات اللغوية.
٤. مرحلة تنفيذ المعجم.
٥. مرحلة اختبار التطبيقات والاستخدامات المستقبلية للمعجم المقترح.
٦. مرحلة الخروج بالنتائج والمؤشرات المبدئية تمهيدًا لما يتلوها من مراحل.

وقدمت بدرية العنزي (٢٠١٥) كذلك رسالة دكتوراه حول بناء نموذجٍ معجمي للمتلازمات اللغوية في العربية المعاصرة في ضوء مدونةٍ لغويةٍ. حيث قامت الباحثة

بجمع المتلازمات اللغوية من المعاجم العربية المعاصرة؛ والمتمثلة في ثلاثة معاجم مختارةٍ، هي: المعجم الوسيط، ومعجم العربية المعاصرة لأحمد مختار عمر، ومعجم العربية المعاصرة المكتوبة لهانزفير. وقد تم جمع هذه المتلازمات بناءً على معاييرٍ محددةٍ وضعتها الباحثة، هي: ضيق المدى، والتوافق، والتواتر. ثم قارنت الباحثة هذه المتلازمات اللغوية بما في المدونة العربية التابعة لمدينة الملك عبد العزيز من خلال المقياس الإحصائي «Log Dice» الذي يتيح في نتائجه العلاقة بين الكلمة والمصاحب لها، وتكرارهما معًا من حيث المتوقع والملحوظ. وكان الهدف العمل على تجسير الفجوة بين المعجم اللغوي والمعجم الاستعمالي (المدونة).

وقد تم في هذه الدراسة اختيار مئة متلازمةٍ لغويةٍ الأكثر ورودًا في المدونة، ووردت في المعاجم الثلاثة، ثم مقارنة معطياتها بمعطيات المدونة حسب المقياس الإحصائي. ثم تقسيمها حسب المجالات الدينية، والسياسية، والاجتماعية، والاقتصادية، والثقافية. ثم دراسة معطيات المدونة العربية من المتلازمات دراسةً تركيبيةً دلاليةً، وإعادة توزيعها حسب نظرية الحقول الدلالية. ومن ثم وضعت الباحثة نموذجًا معجميًا مقترحًا لمعجمٍ عامٍ شاملٍ للمتلازمات اللغوية في العربية المعاصرة.

في الترجمة

نود أن نشير بدايةً إلى إمكانية التحقق من كيفية استثمار المدونات اللغوية في الترجمة بوجهٍ عامٍ ضمن تناولها تفصيلًا في الفصل الثاني.

وتُعد دراسة ماثيو جيدير (Guidere, 2002) من أوئل الدراسات حول كيفية استخدام المدونات اللغوية العربية في الترجمة. حيث أشار الباحث في هذه الدراسة إلى أنه يمكن منهجيًا الجمع بين المقاربة اللغوية والمقاربة الإحصائية لضبط ورفع كفاءة معالجة المدونات اللغوية ثنائية اللغة بما يتسق مع متطلبات الترجمة الآلية.

كما أوضح الباحث أن ذلك يستلزم مطلبين مبدئيين رئيسين، هما:

١. تحديد وتكوين وحدات الترجمة (أي الكلمات، والجمل، والفقرات، والفصول).
٢. استخدام معجمٍ ثنائي اللغة (إنجليزي-عربي، أو فرنسي-عربي).

كما أكد الباحث على أنه ليس هناك حاجةٌ لوَسْم الكلمات نحويًا أو صرفيًا داخل المدونتين اللغويتين المستخدمتين. بل إن هذا يرجع إلى البرامج الآلية في إيجاد المقابلات اللغوية من خلال مقارنة المدونتين اللغويتين ذواتي العلاقة الترجمية فيما بينهما.

وفي الوقت نفسه، شدد الباحث على ضرورة الانتباه إلى ثلاثة أمورٍ، هي:

١. جودة وحجم المعجم ثنائي اللغة المستخدَم. فمن الممكن أن يكون هذا القاموس مبدئيًا جدًّا من ناحية المعلومات النحوية التي يقدمها، غير أنه يتعين أن يكون قادرًا على دمج الكلمات المجهولة الموجودة في المدونة اللغوية ثنائية اللغة المستخدَمة.
٢. نوع البيانات المستخدَمة. فمن الممكن أيضًا أن تُحدِث النصوص ثنائية اللغة المرصوفة المستخدَمة مشكلةً إذا ما كانت جودة المدونة اللغوية ضعيفةً، أو لم تُضبط جيدًا بواسطة أحد الخبراء.
٣. دقة النظام الآلي المستخدَم، وجودة الترجمة المعتمِدة على حجم البيانات التجريبية المتاحة، وإحكام التزامن في المدونة اللغوية.

توالت بعد ذلك الجهود والإسهامات العلمية في مجال الترجمة القائمة على المدونات اللغوية العربية. وكان من بينها دراسة المهنا (Al-Muhanna, 2003) حينما اعتمد على مدونةٍ لغويةٍ للأسماء والصفات المُرَكَبَة في المصطلحات الإنجليزية العلمية والتكنولوجية؛ وذلك في محاولةٍ لنقلها إلى العربية. كما اعتمدت رسالة دكتوراه إزويني (Izwaini, 2004) على مدونةٍ لغويةٍ للمفردات الإنجليزية في تكنولوجيا المعلومات، وترجماتها إلى العربية والسويدية.

في دراسة التوجهات الفكرية (الأيديولوجيا)

ربما تعود أول دراسةٍ عربيةٍ في تحليل التوجهات الفكرية، أو بالأحرى تحليل الخطاب النقدي، اعتمادًا على المدونات اللغوية إلى دراسة الشمري والمحمود (٢٠١٦). ذلك حينما قام الباحثان بتحليل الخطاب النقدي لوعاء الأخبار اعتمادًا على مدونةٍ لغويةٍ من قناتي العربية والجزيرة حول أخبار الحرب على غزة في عام ٢٠١٤. وبلغ عدد كلمات المدونة التي تم تحليلها ٥٢٨٦٤١ كلمةً، موزعةً على ١٥٢٤ نصًا إخباريًا. وتمثل نصوص قناة العربية منها ما نسبته ٣١٪، أما نصوص قناة الجزيرة فتمثل ما نسبته ٦٩٪ من حجم مدونة الدراسة.

وفي سياق تحليلهما، قام الباحثان بتحليلٍ كمي قائمٍ على استخراج قوائم تردد الكلمات، والكلمات المفتاحية لنصوص كل قناةٍ، مع المقارنة بينهما، وتمثيل كشافات النصوص، وتحليل المتلازمات اللغوية للكلمات الأكثر ترددًا، وللكلمات المفتاحية أيضًا. وأشارت نتائج الدراسة إلى وجود سماتٍ لغويةٍ تتسم بها كل قناةٍ، وتميزها عن الأخرى؛ تمثلت هذه السمات المميزة في نوعية الكلمات المستعملة في وصف الأحداث وصياغة الأخبار. إذ تميل قناة العربية إلى الرؤية الرسمية الدولية، مع انتقاء كلماتٍ أكثر عموميةٍ. على عكس قناة الجزيرة التي تميل إلى الوصف والسرد التفصيلي للوقائع وما يحيط بها من ملابساتٍ، وانتقاء كلماتٍ تتضمن معانٍ ثقافيةٍ وتاريخيةٍ ونقديةٍ في سياق وصفٍ دقيقٍ لتفاصيل الحرب على غزة.

وفي العام نفسه، قدم سلطان المجيول (٢٠١٦) دراسةً لخطاب الصحف العربية، والعمل على تحليل أنماط النسوية واللانسوية؛ أي الخطاب المناهض للأنثوية، وذلك المؤيد لها (الأصوات الإيجابية والأصوات السلبية الكمية وغيرهما). ويبدو أن المجيول قام بتطوير أو ربما إعادة نشر هذه الدراسة بالإنجليزية بعد ذلك بعامٍ لتعميم الفائدة (Almujaiwel, 2017).

واستخدمت الدراسة الصحف العربية المتضمنة في المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية (KACST) King Abdulaziz City for Science and Technology Arabic Corpus . وذلك بالاعتماد على مناهج بايبر وكونر وأبتون Biber, Connor and Upton، المصطلح عليها بـ «مناهج بي سي يو BCU approaches».

وتطرقت الدراسة إلى تحليل الصورة الأولية للمنهجية المعتمِدة على مناهج BCU حول موضوعات التحليل: الزواج، والطلاق، والخلع، والحضانة، والعنف، والتعليم، والتمييز، والحجاب، والرياضة، والسفر، وولاية الرجل، وقيادة السيارة، ونزع ولاية الرجل. حيث جُمعت الأنماط الخطابية للنسوية واللانسوية وفقًا لكل موضوعٍ مع الإشارة إلى حالات الأنماط الخطابية دون سياقات الفاعلين فيها، وخلفياتهم الثقافية، ونوعية ما يتمتعون به من سلطةٍ اجتماعيةٍ. وأرجع المجيول السبب في عدم الخوض في نوعيات الفاعلين إلى أهمية كشف الحالات دون الفاعلين، والوقوف على الحالات دون التأثر أو التأثير بنوعيات الفاعلين. كما أن بعض الموضوعات قليلة التكشيف في مدونة الدراسة، نحو: قيادة السيارة، والوصي، والتحرش، ونزع ولاية الرجل متعلقةٌ بالمجتمع السعودي دون المجتمعات العربية الأخرى.

وعلى الرغم من وجود تكراراتٍ لبعض الموضوعات المحددة بمحمل النسوية تفوق عدد تكرارات الموضوع نفسه الذي حمل اللانسوية، وعلى الرغم من وجود العكس أيضًا لبعض الموضوعات، إلا أن المجمل في النتائج الضخمة بجمعها معًا تشير إلى رجحان الإيجابية على السلبية بمعدلٍ بسيطٍ.

دراسات استخدام المدونات اللغوية العربية في استرجاع المعلومات

نود قبل الخوض في استعراض هذه الدراسات أن نشير باديء ذي بدءٍ إلى إمكانية التحقق من الإفادة من المدونات اللغوية في علم المعلومات بوجهٍ عامٍ ضمن تناولها تفصيلًا في الفصل الثاني.

وربما تعود أول دراسةٍ تستخدم مدونةً لغويةً عربيةً في شكلٍ إلكترونيٍ إلى عام ١٩٩٤، ذلك حينما اعتمد الخراشي Alkharashi وإيفينز Evens على مدونةٍ لغويةٍ قوامها ٣٥٥ تسجيلةً ببليوجرافية عربيةً تغطي علوم الحاسب الآلي؛ لإجراء دراسةٍ لتقييم نظام مايكرو-إيرس Micro-AIRS، وهو نظام لاسترجاع المعلومات العربية تم تصميمه كونه نظامًا تجريبيًا لدراسة عمليات التكشيف والاسترجاع للبيانات الببليوجرافية العربية. وفي تجربتهما استعانا بستين استفسارًا صاغها مجموعةٌ من الطلاب العرب المتخصصين في علوم الحاسب الآلي.

وقد وجد الباحثان أنه من بين خمسين استفسارًا هناك تسعةٌ وعشرون استفسارًا لها وثيقةٌ واحدةٌ أو أكثر صالحةٌ للإجابة عليها. وقد توصلت التجربة إلى أن استخدام جذور الكلمات العربية كونها مداخل كشفيةً يعطي نتائج أفضل من استخدام الكلمات في كامل شكلها، وكذلك يعطي نتائج أفضل من الاسترجاع بالاعتماد على تجريد الكلمات stemming.

كما استخدم حميدي Hmeidi (١٩٩٥) في تجربته مدونةً لغويةً قوامها ٢٤٢ مستخلصًا عربيًا؛ بهدف المقارنة بين طرق التكشيف اليدوي وطرق التكشيف الآلي باستخدام الجذوع والجذور والكلمات في كامل شكلها كونها مداخل كشفيةً. وفي سبيل ذلك قام الباحث بحصر كافة الكلمات الواردة في كل وثيقةٍ، ثم رُتبت الكلمات تنازليًا وفقًا لمدى تكرار ترددها، بحيث تُستخدم الكلمات التي وقع معدل تكرار ترددها وفقًا لحدٍ معينٍ كونها مداخل كشفيةً. وفي الوقت نفسه صمم الباحث نظامًا آليًا للتكشيف لتنفيذ تجربته، بحيث يمكنه اختبار عمليات التكشيف والاسترجاع على الوثائق العربية.

وقد أوضحت النتائج على جدوى التكشيف الآلي، وأنه يعطي نتائج أفضل من التكشيف اليدوي عند استخدام الكلمات، والجذوع، والجذور، كونها مداخل كشفيةً. كما كانت النتائج المسترجعة بالتكشيف الآلي الذي يستخدم العبارات أفضل من التكشيف اليدوي باستخدام الكلمات، والجذوع، والجذور. وكانت النتائج المسترجعة بالتكشيف الآلي الذي يستخدم الجذور كونها مداخل كشفيةً أفضل من استخدام الجذوع، وهي النتيجة نفسها التي أعطاها التكشيف اليدوي أيضًا.

وفي عام ٢٠٠١ أصدر ائتلاف البيانات اللغوية Linguistic Data Consortium (LDC) (Gey, Oard & Douglas, 2001) مدونة الأخبار العربية Arabic Newswire. وقد ضمت هذه المدونة ٧٦ مليون هيكل كلمةٍ token و ٦٦٦٠٩٤ كلمةً فريدةً type من النصوص الإخبارية المنشورة بوكالة الأنباء الفرنسية AFP للأخبار العربية بين عامي ١٩٩٤ و ٢٠٠٠. وقد تم تشفير الوثائق بصيغة الشفرة الموحدة Unicode. وقد اعتمدت على هذه المدونة إحدى دراسات مؤتمر تِرِك TREC في تقييم الاسترجاع في اللغة العربية، مُستخدِمةً في ذلك خمسةً وعشرين استفسار بحثٍ باللغة الإنجليزية، مع ترجماتها إلى العربية والفرنسية. ثم طُلب من مجموعةٍ من الـمُحَكمين، قوامها عشر مجموعاتٍ من مؤسساتٍ مختلفةٍ، إصدار قرارات الصلاحية على الوثائق المسترجعة بأحد خيارين؛ «نعم» أو «لا». وبعد توفير هذه المقومات الأساس للتقييم، تم تنفيذ أربعٍ وعشرين عمليةً للاسترجاع الآلي متعدد اللغات باستخدام الاستفسارات الإنجليزية، وثلاث عملياتٍ للاسترجاع الآلي متعدد اللغات باستخدام الاستفسارات الفرنسية، وتسع عشرة عمليةً للاسترجاع الآلي أحادي اللغة باستخدام الاستفسارات العربية، وعمليةٍ واحدةٍ للاسترجاع اليدوي باستخدام الاستفسارات الإنجليزية، وعمليةٍ واحدةٍ للاسترجاع اليدوي باستخدام الاستفسارات العربية. وقد بلغ متوسط عدد الوثائق الصالحة المسترجعة ١٦٥ وثيقةً للاستفسار الواحد.

وفي عام ٢٠٠١ أيضًا قام كريم درويش وزملاؤه (Darwish & et al.) بإنشاء مدونةٍ لغويةٍ أسموها زاد Zad، والتي قامت على توفيرها دار نشر الآريب Al-Areeb Electronic Publishers. وقد احتوت هذه المدونة على أربعة آلاف وثيقةٍ مستخلصةٍ من كتابات ابن القيم في القرن الثالث عشر، وتغطي فروع التاريخ، والفقه، والروحانيات، والسلوكيات. واستخدم الباحثون ٢٥ استفسارًا باللغة العربية للحكم

على صلاحية نصوص مدونة زاد، مع ترجمتها إلى اللغة الإنجليزية؛ وذلك بهدف معالجة المصطلحات الكشفية العربية، والتحقق من الخصائص الصرفية المستخدمة في بحث الإنتاج الفكري العربي، والنقحرة باستخدام حروفٍ إنجليزيةٍ. وقد أظهرت النتائج نجاح تقنية الترجمة المستخدَمة، كما تبين أن البحث في النصوص العربية باستخدام الجذور يعطي نتائج أفضل على المستوى أحادي اللغة.

وخلال الدورة الخامسة والعشرين لمؤتمر مجموعة الاهتمام الخاص باسترجاع المعلومات Special Interest Group on Information Retrieval (SIGIR) اعتمدت إحدى الدراسات (Xu, Fraser & Weischedel, 2002) على المدونة اللغوية العربية لوكالة الأنباء الفرنسية AFP_ARAB لمعالجة اللغة الطبيعية من خلال تقييم بعض استراتيجيات البحث المستخدمة في الاسترجاع أحادي اللغة والاسترجاع متعدد اللغات في اللغة العربية. حيث قامت الدراسة باختبار تأثير تقنين الهجاء spelling normalization وتجريد الكلمات stemming على الاسترجاع متعدد اللغات، استخدمت فيه استفساراتٌ إنجليزيةٌ لاسترجاع الوثائق العربية. وقد أوضحت الدراسة أن هناك بعض الاختلافات في الإملاء العربي، فحرف الياء قد لا يُنقط بدلًا من تسميته ألفًا مقصورةً في نهاية الكثير من الكلمات. ولذا استخدمت الدراسة برنامجًا لتجريد الكلمات يُدعى مُجرِد بَكْوالتر Buckwalter stemmer لتقنين الهجاء عن طريق تصحيح نهايات الكلمات. فيما تغلبت الدراسة على مشكلة الأشكال المختلفة لحرف الألف (أ، إ، ا) عن طريق اختزالها جميعًا في حرف الألف بدون الهمزة (ا). كما اعتمدت الدراسة على مُجرِد بَكْوالتر لرصد كافة الصدور (السوابق) والكواسع (اللواحق) والجذوع واحتمالات تكوِّنها.

وقد توصلت نتائج هذه الدراسة إلى أن تقنين الهجاء وتجريد الكلمات من الممكن أن يُحسِنا الاسترجاع في اللغة العربية على المستوى أحادي اللغة، بينما كان تأثيرهما ضعيفًا في الاسترجاع متعدد اللغات. كما أن تجريد الكلمات، الذي تم على المدونة اللغوية لوكالة الأنباء الفرنسية موضوع الاختبار، لم يُحسِن من كفاءة الاسترجاع بالقدر المأمول.

دراسات الإفادة من المدونات اللغوية في صناعة المكانز

تعد دراسة الثبيتي (٢٠٠٧) من أوائل الدراسات التي اهتمت باستخراج المصطلحات من النصوص اعتمادًا على المدونات اللغوية العربية. حيث اقترحت الدراسة طريقةً إحصائيةً بسيطةً لاستخلاص المصطلحات المتخصصة المكوَّنة من كلمةٍ واحدةٍ اعتمادًا على المدونات اللغوية. وتعتمد هذه الطريقة على حساب «معامل الغرابة» الذي قيس مدى ابتعاد استخدام الكلمة المفردة في إحدى المدونات اللغوية المتخصصة في أحد المجالات، عنها في مدونةٍ لغويةٍ أخرى في أحد المجالات العامة، أو تستخدم اللغة اليومية. وتمثلت المدونة المتخصصة في هذه الدراسة في نصوصٍ في مجال الذكاء الصناعي، ونصوصٍ في مجال الفيزياء. وكان قوام هذه المدونة اللغوية ١٣٢٨٠ هيكل كلمةٍ. بينما تمثلت المدونة اللغوية الأخرى التي تستخدم اللغة اليومية في نصوصٍ تم جمعها من الإنترنت من مواقع سعودية (مواقع صحفٍ، ومجلاتٍ، وكتبٍ، ومطوياتٍ، ومواقع شخصية). وكان قوام هذه المدونة ١٠٣٥٩٩ هيكل كلمةٍ.

وتم تقييم النتائج المسترجعة بواسطة متخصصين في المجالات التي كانت محور الاهتمام والدراسة؛ وهي الذكاء الصناعي، والفيزياء. وأظهرت نتائج الدراسة أن الطريقة المقترحة تقدم مؤشراتٍ مشجعةً في المجالات التي تمت دراستها.

وبعد دراسة الثبيتي بحوالي أحد عشر عامًا، قدمت زايدي، وزملاؤها (٢٠١١) دراسةً أخرى اهتمت باستخراج المصطلحات من النصوص اعتمادًا على المدونات اللغوية العربية. إذ هدفت هذه الدراسة إلى اقتراح آليةٍ لاستخراج المصطلحات من المدونات اللغوية العربية؛ من خلال تقسيم المصطلحات إلى مصطلحاتٍ بسيطةٍ (كلمةٍ واحدةٍ)، ومصطلحاتٍ مركبةٍ على شكل متلازماتٍ لغويةٍ (كلمتين أو أكثر). كما تضمن الاقتراح مقاربةً إحصائيةً للحصول على المصطلحات البسيطة، ومقاربةً لسانيةً لاستخراج المصطلحات المركبة، والتي تخضع للنماذج التالية:

١. اسمٌ-اسمٌ.
٢. اسمٌ-صفةٌ.
٣. فعلٌ-اسمٌ.
٤. اسمٌ-حرفٌ-اسمٌ.

حيث قام الباحثون الأربعة بدايةً بتجريب هذه الطريقة على مدونة الهلال للقرآن الكريم؛ بغرض استخراج المصطلحات، ومن ثم بناء أنطولوجيا بهذه المصطلحات، بحيث يمكن استثمار هذه الأنطولوجيا في أي مجالٍ آخر ، وتوظيفها في تطبيقاتٍ مختلفةٍ، مثل: الترجمة الآلية، والتكشيف الآلي، وتحسين كفاءة البحث على الشبكة العنكبوتية. وقد استخدم الباحثون لتحقيق هدف الدراسة أداة تسمى «جيت Gate» للتحقق من المتلازمات اللغوية في نصوص القرآن الكريم.

وفي عام ٢٠١٤ اقترح ربحي بركة ومنار فياض (Baraka & Fayyad) طريقةً آليةً لاستخراج المصطلحات المتخصصة في مجالٍ معينٍ من مجموعة نصوصٍ عربيةٍ. حيث اعتمد الباحثان في طريقتهما المقترحة على أساليب لغويةٍ وإحصائيةٍ لاستخراج هذه المصطلحات وإيعازها إلى مجالها، مستخدمين مدونةً لغويةً عربيةً مقسمةً إلى عشرة مجالاتٍ؛ وهي المدونة اللغوية أوساك OSAC، التي تحوي نصوص الأخبار العربية المنشورة بموقعي سي إن إن CNN و بي بي سي BBC. واعتمدت طريقة المعالجة على التجريد الخفيف للكلمات light stemming، ومن ثم إمكانية استخراج المصطلحات المرشحة للتضمين. ثم تم تقييم هذه المصطلحات المرشحة وفقًا لوزنها؛ بحيث يتم قبول المصطلحات ذات الوزن الأكبر، لتتكون في النهاية مصفوفةٌ بالمصطلحات. ولاختبار كفاءة الطريقة المقترحة استخدم الباحثان هذه المصفوفة في تكشيف بعض الوثائق للتأكد من ارتباط هذه الوثائق بالمصطلحات المكشفة. وأوضحت النتائج نجاح الطريقة المقترحة بنسبة تحقيق تجاوزت ٩٠٪.

وفي العام نفسه، قدم الثبيتي وزملاؤه (Al-Thubaity & et al., 2014) دراسةً اقترحوا من خلالها طريقتين لاستخراج المصطلحات المكوَّنة من كلمةٍ واحدةٍ، وتلك المكوَّنة من ثلاثة مصطلحاتٍ. وتعتمد الطريقة الأولى على حقيقة أن الكلمات التي يتكرر ترددها واستخدامها في إحدى المدونات اللغوية المتخصصة في مجالٍ ما، يمكن أن تزداد احتمالية الاعتداد بها بصفتها مصطلحاتٍ تعبر عن هذا المجال. بحيث يمكن اعتبار المفردات المكوَّنة من كلمةٍ واحدةٍ، وتلك المكوَّنة من كلمتين، وهذه المكوَّنة من ثلاث كلماتٍ – مصطلحاتٍ دالةً على المجال.

بينما تعتمد الطريقة الثانية على حقيقة أن المصطلحات الدالة على تخصصٍ معينٍ، سواءٌ المكوَّنة من كلمةٍ واحدةٍ أو تلك المركبة، ترتبط فيما بينها ببعض فئات الكلمات

الأخرى، كحروف الجر، والمحددات، وأدوات العطف، وعلامات الإملاء (علامات الترقيم، والأعداد، والعملات، وغيرها من الرموز).

ولتحقيق هدف الدراسة استعان الثبيتي وزملاؤه بمدونةٍ لغويةٍ لنصوص علم اللغة التطبيقي. وأوضحت النتائج أن الطريقتين المقترحتين لاستخراج المصطلحات من المدونات اللغوية صالحتان للتطبيق على المجالات الأخرى. الأمر الذي يفيد كثيرًا في بناء المعاجم والمكانز المتخصصة.

دراسات استخدام المدونات اللغوية في المكتبات

من الدراسات الرائدة في مجال المكتبات والمعلومات حول استخدام المدونات اللغوية في تكشيف الوثائق، تلك الدراسة التي قام بها الباحث التايواني تشن Chen (١٩٩٩) بهدف التحقق من موضوعات الوثائق. حيث اقترح الباحث نموذجًا لمساعدة المستفيدين على استرجاع الوثائق على نحوٍ أكثر فعاليةً. وقد اعتمد هذا النموذج على أربعة معايير رئيسةٍ، هي: أهمية الكلمات، وتردد الكلمات، والمصاحبة بين الكلمات، والمسافة بين الكلمات. وقد طُبق هذا النموذج على مجموعةٍ من النصوص الصينية المتاحة من خلال مدونة سينيكا Sinica Corpus المكونة من 5 مليون كلمةٍ صينيةٍ مُحْشوَّةٍ annotated وموَّسَمَةٍ tagged. وإضافةً إلى ذلك فإن كل نصٍ داخل المدونة اللغوية تمت فهرسته في ضوء خمسة حقولٍ للوصف، هي: النوع الأدبي، والأسلوب، والحالة، والموضوع، والمصدر. وقد انصب الجانب العملي في الدراسة على قياس أداء النموذج المقترح مقارنةً بأداء المستفيدين. وتوصلت الدراسة إلى أن النموذج المقترح في تحليل المدونة اللغوية يوفر وقت وجهد المكتبات في تكشيف الوثائق.

وفي عام ٢٠٠٢ أجرى باحث من الجامعة العبرية (Drori) دراسةً حول التحقق من موضوعات الوثائق المتاحة من خلال المكتبات الرقمية بالاعتماد على مدونةٍ لغويةٍ قوامها ٢٠٠ مقالةٍ علميةٍ بنصوصها الكاملة في موضوعين رئيسين؛ هما: الجغرافيا، ودراسات الأسرة. واستخدمت الدراسة برنامج TextAnalysis في تحليلها للمدونة اللغوية. ويقوم هذا البرنامج بتحليل النصوص وفقًا لتكرار ترددها، ومن ثم يمكن للمستفيدين التحقق من موضوعات هذه النصوص. وقد أوضحت النتائج أن هذا البرنامج يعمل بكفاءةٍ ودقةٍ عاليتين في المساعدة في تحديد الكلمات المفتاحية لمحتويات المكتبات الرقمية. الأمر

الذي من شأنه أن يدعم عمل المسؤولين عن خدمات البحث بالمكتبات الرقمية في تعيين الكلمات المفتاحية للمواد التي تقتنيها بطريقةٍ آليةٍ عالية الكفاءة.

ومن التقارير الرائدة كذلك حول أوجه الإفادة من المدونات اللغوية في المكتبات ذلك التقرير الفني (MacMullen, 2003) الذي نشرته مدرسة المعلومات وعلم المكتبات بجامعة نورث كارولاينا بشابل هيل University of North Caroline at Chapel Hill في مارس عام ٢٠٠٣ حول متطلبات التعريف بالمدونات اللغوية ومعايير تصميمها من أجل تيسير سبل البحث في هذا المجال الناشيء في تخصص المكتبات. وكان من بين أهداف هذا التقرير إبراز أهمية المدونات اللغوية في التحقق من مدى صلاحية، ودقة، وفعالية أدوات البحث، والنظم الآلية للاسترجاع. وقد بدأ التقرير بتناول استخدامات المدونات اللغوية ومدى الاحتياج إليها في المكتبات. ثم ساق التقرير أمثلةً لاستثمار المدونات اللغوية في عددٍ من المجالات؛ كعلم المعلومات، والمعلوماتية الحيوية، واللسانيات، والعلاقات البينية لهذه المجالات، والبرمجيات المستخدَمة. ثم أوضح التقرير الأسس العلمية لتصميم المدونات اللغوية، وهي: التمثيل representativeness، وتحديد واختيار العينة والإحصاء sampling، والاستنساخ، واكتشاف الأخطاء، والتقنين normalization. ثم قدم التقرير مجموعةً من المعايير التي يتعين الالتزام بها عند تصميم المدونات اللغوية بوجهٍ عامٍ. ثم اختتم التقرير عرضه بالتأكيد على أهمية وجدوى المدونات اللغوية في المكتبات، سواءٌ للنهوض بدراساتٍ وأبحاثٍ في المجال، أو باستثمارها في العمليات الفنية، من تكشيفٍ، واستخلاصٍ، وتحليلٍ موضوعيٍ... إلخ.

كما قدم الباحث السلوفيني "كانك" Kanič (2013) ورقةً بحثيةً خلال المؤتمر العلمي الدولي للمدونات اللغوية اقترح فيها مشروعًا لبناء أداةٍ لتنظيم وتقنين مصطلحات المكتبات المتاحة باللغة السلوفينية اعتمادًا على اختزان وتقييم استخدام النصوص العلمية والفنية في المجال. وفي هذا السياق قام الباحث ببناء مدونةٍ لغويةٍ تُيسر عمل موفوضية مصطلحات المكتبات The Commission on Library Terminology وفقًا للإطار الذي أقرته جمعية المكتبات السلوفينية. وتكونت هذه المدونة اللغوية من ١٠٣٠٠ مستخلصٍ شكلت نحو نصف مليون كلمةٍ من النصوص المنشورة قبل عام ١٩٩٩، ما بين رسائل ماجستير، وأطروحات دكتوراه، ومنفرداتٍ، ومقالات دورياتٍ.

وقد بلغ إجمالي ما تم جمعه في هذه المدونة ٦٢٥ مادةً. الأمر الذي من شأنه أن يرفع من أداء اختصاصيي المكتبات، وطلاب المكتبات وعلم المعلومات، إضافةً إلى المعجميين. وقد استخدم الباحث برنامج نيفا NEVA في تحليل المدونة اللغوية.

وخلال ورشة العمل الدولية الأولى للمكتبات الرقمية في علم الموسيقى First International Digital Libraries for Musicology Workshop في عام ٢٠١٤، قام باحثان من مدرسة المكتبات وعلم المعلومات بجامعة إلينوي بإربانا شامبين University of Illinois at Urbana-Champaign، وصاحبهما في ذلك باحثٌ آخر من مكتبة الجامعة (Downie; Dougan & Bhattacharyya, 2014)، بدراسةٍ ببليومتريةٍ لمجموعات مكتبة هاثي تراست الرقمية HathiTrust Digital Library (HTDL) شملت توزيعاتٍ تكراريةً بالمساهمين بالمواد داخل هذه المكتبة، وبلغات المجموعات المقتناه بالمكتبة، وبالموضوعات ورؤوس الموضوعات التي تغطيها هذه المجموعات، وبتواريخ نشرها، والأنواع الأدبية لها. وفي القسم الأخير من الدراسة قام الباحثون ببناء مدونةٍ لغويةٍ لمقتنيات هذه المكتبة من خلال عمل مسحٍ ضوئي لكافة المواد المكتوبة المتاحة، وعلى النحو الذي يُمَّكِن من البحث في النصوص الكاملة لها. وبذلك استطاع الباحثون التحقق من الكلمات المفتاحية لكل مادةٍ، ومن ثم الاستفادة منها في صياغة رؤوس الموضوعات المناسبة. وفي النهاية أكد الباحثون على جدوى تضمين تكشيف الكلمات المفتاحية القائم على مدونةٍ لغويةٍ في تيسير استرجاع مقتنيات المكتبات.

وإجمالًا لما سبق، وبالرجوع إلى مجالات الإفادة من المدونات اللغوية في الفصل الثاني من هذا الكتاب، فإننا نلاحظ أنه لا تزال هناك مجالاتٌ عدةٌ لم تُستثمر فيها المدونات اللغوية العربية بعد. كما هو الحال في مجالات صناعة المكانز، والمعلوماتية الحيوية، والمعلوماتية الجنائية.

ونلاحظ أيضًا أن الدراسات التي تناولت المدونات اللغوية في البيئة العربية جاءت معظمها على يد باحثين عرب لكن باللغة الإنجليزية. وقد اقتصرت معظم هذه الدراسات على كيفية استثمار المدونات اللغوية في اختبار وتقييم نظم استرجاع المعلومات ومعالجة اللغة العربية.

الخلاصة

انصب هذا الفصل على الإنتاج الفكري الذي تناول المدونات اللغوية؛ سواءٌ من حيث الإنشاء والتصميم والإتاحة، أو من حيث الاستخدام المنهجي. وكان الهدف من ذلك تقديم مراجعةٍ علميةٍ انتقائيةٍ لهذا الإنتاج حول الموضوع، مع التركيز على الإسهامات العربية قدر الإمكان. وفي هذا السياق تم تناول تلك الدراسات المهتمة بإنشاء وإتاحة المدونات اللغوية العربية، ودراسات الإفادة منها في علم اللغة التطبيقي (النحو والدلالة، وعلم اللغة الاجتماعي، وصناعة المعاجم، والترجمة، وتحليل التوجهات الفكرية)، ودراسات الاستخدام في استرجاع المعلومات، وبناء المكانز، ودراسات استثمارها في المكتبات.

قائمة ببليوجرافية

أولًا: المراجع العربية

ابن منظور، جمال الدين أبو الفضل محمد بن مكرم. (١٩٩٣). مادة «ذخر». في: لسان العرب (مج ٤). بيروت: دار صادر. استرجع من.http://shamela.ws/browse.php/book-1687#page-2262

بوشحدان، الشريف. (٢٠١٠). الأستاذ عبد الرحمن الحاج صالح وجهوده العلمية في ترقية استعمال اللغة العربية. *مجلة كلية الآداب والعلوم الإنسانية والاجتماعية*، ٧. ٥٠-٢٥.

الثبيتي، عبد المحسن. (٢٠٠٧). استخدام ذخائر النصوص لاستخلاص المصطلحات المتخصصة. *الندوة الدولية الأولى عن الحاسب واللغة والعربية*. مدينة الملك عبدالعزيز للعلوم والتقنية. الرياض. المملكة العربية السعودية.

الثبيتي، عبد المحسن. (٢٠١٦). الكلمات المُمَيَّزَةُ للمدونات اللغوية: قضايا تقنية. فصل من كتاب: *لغويات المدونة الحاسوبية: تطبيقات تحليلية على العربية الطبيعية*. تحرير سلطان المجيول. الرياض: مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية.

حمادة، سلوى. (٢٠١١). نحو منهج عربي مقترح لتصميم المدونات اللغوية. استرجع من http://www.globalarabnetwork.com/science-a-it/2784-2011-04-04-14-49-07

الربيعة، مها سليمان؛ السلمان، عبد الملك سلمان & آتويل، إيريك. الذخيرة اللغوية الفصحى حجر الأساس لدراسة معاني كلمات القرآن الكريم: في ضوء نماذج الدلالة المتوزعة. استرجع من http://ksucorpus.ksu.edu.sa/wp-content/uploads/2014/01/الذخيرة-اللغوية-الفصحى-حجر-الأساس-لدراسة-معاني-كلمات-القرآن-الكريم.pdf

زايدي، صورية؛ عبد العالي، أحمد؛ العسكري، محمد الطيب & الشنيفي، محمد عبد الله. (٢٠١١). استخراج المصطلحات البسيطة والمركبة من النّصوص العربية: تطبيق على النّص القرآني الكريم. Communications of the Arab Computer Society، 41.

الزهيري، نبيل. (٢٠٠٣). قاموس مصطلحات المعلوماتية واللغويات الحسابية: إنجليزي – عربي مع مسارد بالإنجليزية والعربية. بيروت: مكتبة لبنان. ٧٠٤ ص.

الشاذلي، هشري. (١٩٩٨). الضمير: بنيته ودوره في الجملة (رسالة دكتوراه غير منشورة). جامعة منوبة، كلية الآداب.

الشامي، أحمد محمد & حسب الله، سيد. (٢٠١٤). مصطلحات المكتبات والمعلومات والأرشيف. استرجع من http://www.elshami.com

الشمري، عقيل بن حامد & المحمود، محمود بن عبد الله. (٢٠١٦). المعالجة الآلية لوعاء الأخبار: تحليل الخطاب النقدي المعتمد على المدونة الحاسوبية. في كتاب: *لغويات المدونات الحاسوبية: تطبيقات تحليلية على العربية الطبيعية، الرياض:* مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، ١٩٨-٢٥٢.

الصوينع، علي سليمان. (يوليو ١٩٨٧). كشافات النصوص وتطبيقاتها في نصوص القرآن والحديث. *مجلة المكتبات والمعلومات العربية*. ٧(٣). ٥-٥٢.

عبد العالي، أحمد عبد الله. (١٩٩٦). الببليوجرافيا والتكشيف في المكتبات. (سلسلة تبسيط علوم المكتبات، ٤). الكويت: وزارة التربية، إدارة المكتبات. ٢٧ ص.

عبد الهادي، محمد فتحي. (١٩٨٢). التكشيف لأغراض استرجاع المعلومات. جدة: مكتبة العلم. ٢١٣ ص.

______________.(١٩٨٤). مقدمة في علم المعلومات. القاهرة: دار غريب. ٣٢٠ ص.

عبد الهادي، محمد فتحي & زايد، يسرية عبد الحليم. (٢٠٠٠). التكشيف والاستخلاص: المفاهيم، الأساس، التطبيقات. القاهرة: الدار المصرية اللبنانية. ٢٤٤ ص.

العتيق، زايد بن مهلهل. (١٩٩٢). تحليل الأخطاء الدلالية لدى دارسي اللغة العربية من غير الناطقين بها في مادة التعبير الكتابي. (رسالة ماجستير غير منشورة). جامعة الإمام محمد بن سعود الإسلامية.

عرفات، كمال نبهان. (٢٠٠٩). تكشيف نصوص التراث العربي. القاهرة: مكتبة الإمام البخاري للنشر والتوزيع. ٢٥٦ ص.

العصيلي، عبد العزيز بن إبراهيم. (١٩٨٥). الأخطاء الشائعة في الكلام لدى طلاب اللغة العربية الناطقين بلغات أخرى (رسالة ماجستير غير منشورة). جامعة الإمام محمد بن سعود الإسلامية.

عمر، أحمد مختار. (١٩٨٨). البحث اللغوي عند العرب: مع دراسة لقضية التأثير والتأثر. القاهرة: عالم الكتب. ٣٨٣ ص.

عمر، أحمد مختار. (٢٠٠٨). معجم اللغة العربية المعاصرة. ط ١. القاهرة: عالم الكتب. ٣٣٦٨ ص.

العنزي، بدرية. (٢٠١٥). نحو بناء معجم للمتلازمات اللفظية في المعاجم العربية المعاصرة: دراسة تحليلية في ضوء مدونة لغوية. رسالة دكتوراه. جامعة الإمام محمد بن سعود الإسلامية.

غزالة، حسن. (٢٠٠٧). قاموس دار العلم للمتلازمات اللفظية: قاموس شامل إنجليزي-عربي لمعاني الألفاظ وتواردها ودقة استعمالها. بيروت: دار العلم للملايين. ١٥٢٧ ص.

قاسم، حشمت. (٢٠٠٠). مدخل لدراسة التكشيف والاستخلاص. القاهرة: دار غريب. ٣٠٠ ص.

القاموس المتعدد اللغات على الإنترنت Glosbe. استرجع من https://ar.glosbe.com/

قرآني = Qurany. استرجع من http://quranytopics.appspot.com

المالكي، هشام موسى. (٢٠٠٩). إشكاليات تهيئة الذخائر النصية وبنائها حاسوبيًا: اللغتان العربية والصينية نموذجًا. *مجلة علوم اللغة*، ٦٤. استرجع من http:// www.afaaq-edu.com/photo_gallery/5-corpus%20building%20 -%20a%20wen.pdf

المجيول، سلطان. (٢٠١٦). مناهج التهيئة المعجمية في تعليم العربية لغير الناطقين بها. *المؤتمر الدولي الثاني في الاتجاهات الحديثة في تعليم العربية لغة ثانية*. الرياض: معهد اللغويات العربية بجامعة الملك سعود.

ــــــــــــ (٢٠١٦). المعالجة الآلية للصحف العربية: تحليل الأنماط الخطابية بمناهج BCU. في كتاب: *لغويات المدونات الحاسوبية: تطبيقات تحليلية على العربية الطبيعية*. الرياض: مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية. ١٢-٥٦.

المدونة اللغوية العربية العالمية لمكتبة الإسكندرية = International Corpus of Arabic. استرجع من http://www.bibalex.org/ica/ar

ميدار: المشروع المتوسطي لتقنيات اللغات العربية المكتوبة والمنطوقة (مشروع من سلسلة مشروعات نملار). استرجع من http://www.medar.info/Breif_ Arabic/MEDAR _Arabic -brief-June2009.pdf

ثانيًا: المراجع الأجنبية

Abbas, Q. (2012). Building a hierarchical annotated corpus of Urdu: The URDU.KON-TB Treebank. *CICLing 2012*, Part I, LNCS 7181. 66–79. Retrieved from http://ling.uni-konstanz.de/pages/home/pargram_urdu/main/fil es/Abbas2012.pdf

Abbasi, A.; Chen, H. & Salem, A. (2008). Sentiment analysis in multiple languages: feature selection for opinion classification in web forums ACM Trans. *Inf. Syst*., 26. 1–34

Abdelali, A. (2004). Localization in Modern Standard Arabic. *Journal of the American Society for Information Science and Technology*. 55(1), 23-28.

Abdelali, A.; Cowie, J. & Soliman. H. (2005). Building A Modern Standard Arabic Corpus. *Workshop on Computational Modeling of Lexical Acquisition*. The Split Meeting. Croatia, 25th to 28th of July 2005.

AbdelRaouf, A. & et al. (2010). Building a multi-modal Arabic corpus (MMAC). *IJDAR*, 13. 285–302.

Abdul Razak, Z. (2014). Word Usage Variations in Arabic Newspapers: A Corpus Investigation. *GEMA Online® Journal of Language Studies*. 14(2). 29-45.

Abu El-Khair, I. (November 2016). Abu El-Khair Corpus: A Modern Standard Arabic Corpus. *International Journal of Recent Trends in Engineering & Research (IJRTER)*, 2(11). 5-13.

Al-Ansary, S. (2005). Building a Computational Lexicon for Arabic: A corpus-based approach. In M. Alhawary & B. Elabbas (eds.) *Perspectives on Arabic Linguistics*. 173–93.

Alansary, S.; Nagi, M. & Adly, N. (2008). Building an International Corpus of Arabic (ICA): Progress of Compilation Stage.

Alansary, S. & Nagi, M. (2014). The International Corpus of Arabic: Compilation, Analysis and Evaluation. *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, October 25, 2014, Doha, Qatar. 2014 Association for Computational Linguistics. 8–17.

Alfaifi, A.; Atwell, E. & Abuhakema, G. (2013). Error Annotation of the Arabic Learner Corpus: A New Error Tagset. *Lecture Notes in Computer Science*, 8105. 14-22.

Alfaifi, A. & Atwell, E. (2014). Arabic Learner Corpus and Its Potential Role in Teaching Arabic to Non-Native Speakers. *The 7th Biennial IVACS conference*, 19 - 21 Jun 2014. Newcastle, UK.

Al-Kharashi, I. A. & Evens, M. W. (1994, Sept.) Comparing words, stems and roots as index terms in an Arabic information retrieval system. *Journal of American Society for Information Science*, 45(8). 548-560.

Almeman, K. & Lee, M. (2013). Automatic Building of Arabic Multi Dialect Text Corpora by Bootstrapping Dialect Words. *1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, 2013.

Al-Muhanna, A. (2003). Scientific and technological terms transfer into Arabic: A corpus-based study of Arabic noun+noun and noun+adjective compounds, Ph. D. thesis, UMIST.

Almujaiwel, S. & Al-Thubaity, A. (2016). Arabic Corpus Processing Tools for Corpus Linguistics and Language Teaching. *The Globalization of Second Language Acquisition and Teacher Education*, August 4-6, 2016, Fukuoka, Japan.

Almujaiwel, S. (2017). Discursive patterns of anti-feminism and pro-feminism in Arabic newspapers of the KACST corpus. *Discourse & Communication*, 11(5), 441-466.

Al-Sulaiti, L. & Atwell, E. (2006). The design of a corpus of Contemporary Arabic. *International Journal of Corpus Linguistics*, 11(1). 1–36.

Al-Thubaity, A. M., Khan, M., Alotaibi, S., & Alonazi, B. (2014, October). Automatic Arabic term extraction from special domain corpora. In *International Conference on Asian Language Processing (IALP)*, 1-5.

Al-Thubaity, A. (2015). A 700M+ Arabic corpus: KACST Arabic corpus design and construction. *Language Resources and Evaluation*, 49(3), 721-751.

Al-Thubaity, A.; Khan, M.; Al-Mazrua, M. & Al-Mousa, M. (2013, August). New Language Resources for Arabic: Corpus Containing More Than Two Million Words and a Corpus Processing Tool. In *Asian Language Processing (IALP), 2013 International Conference on IEEE.* 67-70.

Anthony, L. (2017). AntConc [Computer Software]. Tokyo, Japan: Waseda University. Available at http://www.laurenceanthony.net/

arabiCorpus: the Arabic corpus for the rest of us. Retrieved from http://arabic orpus.byu.edu/

Aston, G. (1999). Corpus use and learning to translate. *Textus* 12. 289-314. Retrieved from http://www.sslmit.unibo.it/~guy/textus.htm.

Attia, M.; Tounsi, L & Genabith, J. (2010). Automatic Lexical Resource Acquisition for Constructing an LMF-Compatible Lexicon of Modern Standard Arabic. *The NCLT Seminar Series*, DCU, Dublin, Ireland. Retrieved from http://www.attiaspace.com/Publications/Corpus-Driven_Lexicon_of_MS A.pdf

Baker, M. (1995). Corpora in translation studies: an overview and some suggestions for future research. *Target,* 7(2). 223-243.

Baraka, R. & Fayyad, M. (2014). Automatic Domain-Relevant Collocation Extraction from Arabic Corpus. *IUG Journal of Natural and Engineering Studies*, 22(2), 30-44.

Barbu, C.; Evans, R. & Mitkov, R. (2002). A corpus based investigation of morphological disagreement in anaphoric relations. In

Proceedings of the Language Resources and Evaluation Conference.

Benajiba, Y. & Rosso, P. (2007). Towards a measure for Arabic corpora quality. *Proceeding of International Colloquium on Arabic Language Processing, CITALA-2007*. Retrieved from http://www.researchgate.net/publication/22897 2993_Towards_a_measure_for_arabic_corpora_quality/file/9fcfd51421b952e785.pdf

Bennett, G. (2010). Using Corpora in the Language Learning Classroom Corpus Linguistics for Teachers. Michigan: University of Michigan Press. 144 p.

Berber-Sardinha, T. (2000). Comparing corpora with WordSmith Tools: How large must the reference corpus be? *Proceedings of the Workshop on Comparing Corpora 9*, 7-13.

Biber, D. (1994). Representativeness in Corpus Design. *Linguistica Computazionale*, 9. 377-407. Retrieved from http://link.springer.com/chapt er/10.1007%2F978-0-585-35958-8_20

Blecha, J. (2012). Building Specialized Corpora. M.A. Thesis. Masaryk University. Retrieved from http://is.muni.cz/th/179991/ff_m_b1/1799 91_Building_Spe cialized_Corpora.pdf

Borko, H. & Bernier, C. (1978). Indexing concepts and methods. London: Academic press. 271 p.

Bowker, L & Pearson, J. (2002). Introducing corpora and corpus analysis tools. In: *Working with specialized language: a practical guide to using corpora*. London: Routledge. 242 p.

Buckwalter, T. & Parkison, D. (2011). A Frequency Dictionary of Arabic: Core Vocabulary for Learners. New Yok and London: Routledge, Taylor, Francis Group.

Bunt, Harry & Black, Bill. (2000). The ABC of computational pragmatics. Computational pragmatics: Abduction, belief and context, ed. by Harry C. Bunt and William Black. Amsterdam:

John Benjamins.

BYU-BNC: British National Corpus. (2015). Retrieved from http://corpu s.byu.edu/bnc/

Chen, K. (1999). Automatic Identification of Subjects for Textual Documents in Digital Libraries. Los Alamos, NM: Los Alamos National Laboratory.

Chomsky, N. (1965). Aspects of the Theory of Syntax. Cambridge, Mass.: MIT Press. 315 p.

Collins. (2015). Retrieved from http://www.collinsdictionary.com/

———. (2015). Retrieved from http://www.collinsdictionary.com/english-thesaurus

Corpora at Victoria University of Wellington. (2013). Retrieved from http://www.v ictoria.ac.nz/lals/resources/corpora-default

CQPweb v3.2.25 [Computer Software]. Retrieved from https://cqpweb.lancs.ac.uk/

Culpeper, J. (2009). Keyness: Words, part-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics*, 14(1). 29-59.

Darwish, K. et al. (2001). TREC-10 Experiments at University of Maryland CLIR and Video (2001). In: *Proceedings of the Tenth Text REtrieval Conference (TREC-2001)*, NIST Special. Retrieved from http://citeseer.uark.edu:808 0/citeseerx/viewdoc/summary?doi=10.1.1.10.8215

Dash, N. (2015). The art of lexicography. *UNESCO - Encyclopedia of Life Support System (EOLSS)*. Retrieved from http://www.eolss.ne t/sample-chapters/c04/e6-91-16.pdf

Davies, M. (2011-). Google Books (American English) Corpus (155 billion words, 1810-2009). Retrieved from http://googlebooks.byu.edu/.

Dictionary.com. (2015). Retrieved from http://dictionary.reference.com/

Downie, J.; Dougan, K. & Bhattacharyya, S. (2014). The HathiTrust Corpus: A Digital Library for Musicology Research? First International Digital Libraries for Musicology Workshop (DLfM 2014), London, UK. Retrieved from http://www.music-ir.org/mirex/DLfM_proceedings/long_papers/dlfm2014_submission_33.pdf

Drori, O. (2002). Identifying the Subject of Documents in Digital Libraries Automatically Using. *Proceedings of the 3rd International Workshop on New Developments in Digital Libraries*. Retrieved from http://leibniz.cs.huji.ac.il/t r/acc/2002/HUJI-CSE-LTR-2002-40_drori062002b.pdf

Dukes, K.; Atwell, E. & Sharif, A. (2010). Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Retrieved from http://hnk.ffzg.hr/bibl/lrec2010/pdf/278_Paper.pdf

Dukes, K. & Habash, N. (2010). Morphological Annotation of Quranic Arabic. *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, 17-23 May 2010, Valletta, Malta. 2530-2536

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1). 61-74.

Eddakrouri, A. (2016). ba'da and qabla in online news: A corpus-based study. M.A. Thesis. The American University in Cairo, the Department of Applied Linguistics. Retrieved from http://dar.aucegypt.edu/bitstream/handle/10526/48 02/Ayman-%20Thesis.pdf?sequence=1

El-Shishtawy, T. & El-Ghannam, F. (2012). An accurate Arabic root-based lemmatizer for information retrieval purposes. *International Journal of Computer Science Issues (IJCSI)*, 9 (1).

Essa, S. (2013). Corpus-based analysis of three Arabic adversative conjunctions in current Egyptian newspapers. M.A. Thesis. The American University in Cairo, the Department of Applied Linguistics. Retrieved from http://dar.aucegypt.edu/bitstream/handle/10526/3549/Essa%2c%20Shaemaa%28May%2c2013%29.pdf?sequence=3

Feather, J & Sturges, P. (2003). International encyclopedia of information and library. London: Routledge.

Fowler, R. (1987). Notes on Critical Linguistics. In R. Steele and T. Threadgold (eds.) Language Topics: Essays in honour of Michael Halliday vol. II. Amestrdam: Benjamine. 481-492. Reprinted with modifications as: On Critical Linguistics, in Caldas-Coulthard and Coulthard (eds.) 3-14.

Francis, W. & Kucera, H. (1979). Brown Corpus manual. Rhode Island: Brown University. Retrieved from http://clu.uni.no/icame/brown/bcm.html

Garfinkel, S. (2009). Bringing science to digital forensics with standardized forensic corpora. *Digital Investigation: The International Journal of Digital Forensics & Incident Response*, 6. 2-11. Retreived from http://digitalcorpor a.org/wp/wp-content/uploads/2010/03/p2-garfinkel.pdf

Garg, S.; Martinovski, B. & Robinson, S. (2004). Evaluation of Transcription and Annotation tools for a Multi-modal, Multi-party dialogue corpus. In *Proceedings of the LREC*. Retrieved from http://people.ict.usc.edu/~traum/ Papers/tools6.pdf .

Garside, R.; Leech, G. & Sampson, G. (1987). The Computational analysis of English: A Corpus-based approach. London: Longman.

Genome Annotation Assessment Project - GASP1. (2014). Retrieved from http://www.fruitfly.org/GASP1/

Gey, F. & Oard, D. (2001). The TREC-2001 Cross-language information

retrieval track; searching Arabic using English, French or Arabic queries. Retrieved from http://trec.nist.gov/pubs/trec10/

Ghazali, S. & Braham, A. (2001). Dictionary Definitions and Corpus-Based Evidence in Modern Standard Arabic. *Arabic NLP Workshop at ACL/EACL*. Toulouse, France.

Global English Monitor Corpus. Retrieved from http://www.lt-world.org/kb/resou rces-and-tools/language-tools/systems/obj_76622

Goh, G. (2011). Choosing a Reference Corpus for Keyword Calculation. *Linguistic Research,* 28(1). 239-256. Retrieved from http://isli.khu.ac.kr/jour nal/content/data/28_1/13.pdf

Google Books Corpus. Retrieved from http://googlebooks.byu.edu/

Goweder, A. & De Roeck, A. (2001). Assessment of a Significant Arabic Corpus. In: *the Arabic NLP Workshop at ACL/EACL 2001*. Retrieved from http://www.abdelali.net/ref/ACL-EACL%20 2001_goweder.pdf

Grefenstette, G. (1993). Automatic Thesaurus Generation from Raw Text using Knowledge-Poor Techniques. In *Making Sense of words, Ninth Annual Conference of the UW Center for the New OED and Text Research*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.2 1.6422&rep=rep1&type=pdf

————————. (1994). Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers, Boston, USA.

Gries, S. & Berez, A. (2013). Linguistic annotation in/for corpus linguistics. University of California, Santa Barbara and University of Hawai'i at Mānoa.

Guidere, M. (2002) Toward Corpus-Based Machine Translation for Standard Arabic. *Translation Journal*, 6(1). Retrieved from http://www.mt-archive.info/TranslationJ-2002-Guidere.pdf

Haan, P. (1984). Problem-oriented tagging of English corpus data. In *Aarts, J. and Meijs, W. (eds) Corpus Linguistics*. Amsterdam: Rodopi.

Habash , N. & Rambow , O. (2005). Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting of the ACL*, 573-580, Ann Arbor, MI.

Habash, N. & Roth, R. (2009). CATiB: The Columbia Arabic Treebank. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Retrieved from http://www.researchgate.net/profile/Ryan_Roth/publication/228946555_Syntactic_Annotation_in_the_Columbia_Arabic_Treebank/links/0deec517837b2f0209000000.pdf

Hammami, S.; Belguith, L. & Hamadou, A. (2009). Arabic Anaphora Resolution: Corpora Annotation with Coreferential Links. *The International Arab Journal of Information Technology*, 6 (5). 481 – 489.

Hammo, B.; Yagi, S.; Ismail, O. & AbuShariah, M. (2016). Exploring and exploiting a historical corpus for Arabic. *Language Resources and Evaluation*, 50(4), 839-861.

Hanks, P. (2000). Literal and Metaphorical Word Meaning. Toscan Word Center document.

Helsinki Corpus of English Texts. (2011). Retrieved from http://www.helsinki.fi/v arieng/CoRD/corpora/HelsinkiCorpus/

Hmeidi, I. (1995). Design and implementation of automatic word and phrase indexing for information retrieval with Arabic documents. (Doctoral dissertation). Illinois Institute of Technology. 130 p.

Hovy, E. & Lavid, J. (2010). Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *International Journal of Translation*, 22(1), 13-36.

Hunston, S. (2002). Corpora in Applied Linguistics. Cambridge: Cambridge University Press.

Ismail, A. (2015). tab asta'zen ana ba'a: a corpus-based study of

three discourse markers in Egyptian film language. M.A. Thesis. The American University in Cairo, the Department of Applied Linguistics. Retrieved from http://dar.aucegypt.edu/bitstream/handle/10526/4417/My%20MA%20Thesis%20%28Signature%20included%29.pdf?sequence=3

Izwaini, S. (2004). Translation and The Language of Information Technology: A Corpus-based Study of the Vocabulary of Information Technology in English and its Translation into Arabic and Swedish. PhD thesis. University of Manchester

Jan, H. et al. (2004). Prague Arabic Dependency Treebank 1.0. Retrieved from https://catalog.ldc.upenn.edu/ldc2004t23

Jing, Y. & Croft, W. (1994). An Association Thesaurus for Information Retrieval. In: *Bretano, F., Seitz, F.: (eds.): Proceedings of the RIAO'94 Conference*. CIS-CASSIS, Paris, France. 146-160

Jones, K. (1971). Automatic Keyword Classification for Information Retrieval. Butterworths, London, UK (1971).

Kanič, I. (2013). Slovene Specialized Text Corpus of Library and Information Science – an Advanced Lexicographic Tool for Library Terminology Research. *Proceedings of the international scientific conference*, 2013. Retrieved from http://corpora.phil.spbu.ru/Works2013/Kani%C4%8D.pdf

Katz, S. (1996). Distribution of Common Words and Phrases in Text and Language Modelling. *Natural Language Engineering*, 2 (1). 15-59

Khoja, S.; Garside, R. & Knowles, G. A tagset for the morphosyntactic tagging of Arabic. Retrieved from http://archimedes.fas.harvard.edu/mdh/arabic/CL200 1.pdf

Khosrow-Pour, M. (2015). Encyclopedia of Information Science and Technology, Third Edition. Hershey: Information Science Reference.

Kilgarriff, A. & Grefenstette, G. (2003). Web as Corpus. Retrieved from http://www.kilgarriff.co.uk/Publications/2003-KilgGrefenstette-WACIntro.pdf

LE ROPERT. (2015). Retrieved from http://www.lerobert.com/

Ledwith, R. (1992). On the difficulties of applying the results of information retrieval research to aid in the searching of large scientific databases. *Information Processing and Management*, (28)4. 451-55.

Lee, H. & et al. (2013). CoMAGC: a corpus with multi-faceted annotations of gene-cancer relations. *BMC Bioinformatics*, 14. Retrieved from http://www.b iomedcentral.com/1471-2105/14/323

Leech, G. & Wilson, A. (1996). Recommendations for the Morphosyntactic Annotation of Corpora EAGLES Report EAG-TCWG-MAC/R. http://www.il c.pi.cnr.it/EAGLES96/annotate/annotate.html

Leech, G. (1997). Corpus annotation: Linguistic information from computer text corpora, chapter Introducing corpus annotation. London: Longman. 344 p.

Leech, G.; McEnery A. & Wynne, M. (1997). Further levels of annotation. In *Garside, R. G., Leech, G. N. & McEnery, A. M. (eds.). Corpus annotation*. Longman: Longman. 85-101.

Lijffijt, J. (2011). Analyzing word frequencies in large text corpora using inter-arrival times and bootstrapping. *Lecture Notes in Computer Science*, 6912. 341-57

Louw, B. (1993). Irony in the text or insincerity in the writer? the diagnostic potential of semantic Prosodies. in Baker, M. and Tognini-Bonelli, E. (eds.) *Text and Technology: in Honour of John Sinclair*. Amsterdam: John Benjamins. 157-176.

Maamouri, M.; Bies, A.; Jin, H. & Buckwalter, T. (2003) Arabic

Treebank: Part 1 v 2.0. Retrived from https://catalog.ldc.upenn.edu/LDC2003T06

Macmillan Dictionary. (2015). Retrieved from http://www.macmillandictionary.com/

MacMullen, W. (2003). Requirements Definition and Design Criteria for Test Corpora in Information Science. *SILS Technical Report*. Retrieved from https://sils.unc.edu/sites/default/files/general/research/TR-2003-03.pdf

Mansour, M. (June, 2013). The Absence of Arabic Corpus Linguistics: A Call for Creating an Arabic National Corpus. *International Journal of Humanities and Social Science*, 3(12). 81-90.

Margaretha, E. & Lüngen, H. (2014). Building Linguistic Corpora from Wikipedia Articles and Discussions. *Journal of Language Technology and Computational Linguistics*, 29(2). 59-82. Retrieved from http://ww w.jlcl.org/2014_Heft2/3MargarethaLuengen.pdf

McEnery, T. (2013). Corpus linguistics. In R. Mitkov (ed), Handbook of Computational Linguistics, Oxford University Press. 448-463.

————. (2006). Using available corpora. Retreived from http://www.lancaste r.ac.uk/fass/projects/corpus/ZJU/xCBLS/chapters/A07.pdf .

————. (2009). Keywords and moral panics: Mary Whitehouse and media censorship. In *D. Archer (ed.) What's in Word-list? Investigating Word Frequency and Keyword Extraction*. Oxford: Ashgate.

McEnery, T.; Xiao, R. & Tono, Y. (2006). Corpus-Based Language Studies: An advanced resource book. London and New York: Routledge.

McIntyre, D. et al. (2004). Investigating the presentation of speech, thought and writing in spoken British English: a corpus-based approach. *International Journal of Corpus Linguistics*, 28. 49-76.

Mitkov, R.; Belguith, L. & Malgorzata, S. (1998). Multilingual Robust Anaphora Resolution. in *Proceedings of the 3rd International Conference on Empirical Methods in Natural Language Processing*, Grenade. 7-16.

Mohamed, E. & Kubler, S. (2010). Arabic Part of Speech Tagging. *LREC 2010, Seventh International Conference on Language Resources and Evaluatio*n. Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/summaries/384.html

Morin, E. & Jacquemin, C. (1999). Projecting Corpus-Based Semantic Links on a Thesaurus. *ACL '99 proceedings of the 37th annual meeting of the Association for Computational Linguistics*. 389-396

Nicholls, D. (2003). The Cambridge Learner Corpus - error coding and analysis for lexicography and ELT. Retrieved from http://ucrel.lancs.ac.uk/publications/cl 2003/papers/nicholls.pdf

NOW Corpus (News on the Web). Retrieved from http://corpus.byu.edu/now/

O'Keeffe, A. & Farr, F. (2003). Using language corpora in language teacher education: pedagogic, linguistic and cultural insights. *TESOL Quarterly*, 37(3). 389-418.

O'Keeffe, A.; McCarthy, M. & Carter, R. (2007). From corpus to classroom: language use and language teaching. Cambridge: Cambridge University Press. 315 p.

Parkinson, D. (2003). Future Variability: A Corpus Study of Arabic Future Particles. In D. Parkinson & S. Farwaneh (Eds.), *Perspectives on Arabic Linguistics XV: Papers from the Fifteenth Annual Symposium on Arabic Linguistic*s (pp.191–211). Salt Lake City.

————. (2008). Sentence Subject Agreement Variation .In Ibrahim, Z. & Makhlouf, S. A. M. (Ed.). *Linguistics in Age of Globalization: Perspectives on Arabic Language and Teaching* (pp. 67-90). Cairo: American University Cairo Press.

Pearce, M. (2006). The Routledge Dictionary of English Language Studies. New York: Routledge. 211 p.

Pearson ELT: Longman Corpus Network. Retrieved from http://www.pearsonlong man.com/dictionaries/corpus/written-american.html

Querying Arabic Corpora [Website]. Retrieved from http://corpus.leeds.ac.uk/quer y-ar.html

Richard, J. C. & Schmidt, R. (2002). Longman Dictionary of Language Teaching & Applied Linguistics. London: Pearson Education.

Rychly, P. & Kilgarriff, A. (2007). An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. 41-44.

Salton, G. (1971). Experiments in Automatic Thesauri Construction for Information Retrieval. In *Proceedings of the IFIP Congress*, TA-2. Ljubljana, Yoguslavia. 43-49

————. (1971). The SMART Retrieval System: Experiments in automatic document processing. Englewood Cliffs, NJ: Prentice-Hall.

Samy, D. & González-Ledesma, A. (2008). Pragmatic Annotation of Discourse Markers in a Multilingual Parallel Corpus (Arabic-Spanish-English). *LREC*.

Sanderson, M. (1996). The Reuters test collection. In *Leon, Ruben, ed. Proceedings of the Sixteenth Research Colloquium of the British Computer Society Information Retrieval Specialist Group, Drymen, Scotland*, 22-23. London: Taylor Graham. 219-27.

Sawalha, M. & Atwell, E. (2008). Comparative evaluation of Arabic language analyzers and stemmers. In: *Coling 2008: Posters and Demonstrations*. 107-110.

Schneider, J. (2004). Verification of bibliometric methods' applicability for thesaurus construction. (Doctoral dissertation). Aalborg: Department of Information Studies, Royal School of Library and Information Science.

School of Arts, Languages and Cultures: A Representative Corpus of Historical English Registers. Retrieved Feb. 15, 2015, from http://www.alc.manchester. ac.uk/subjects/lel/research/projects/archer/

Scott, M. (1997). PC Analysis of Key Words - and Key Key Words. *System*, 25(2). 233-45.

———. (2009). In search of a bad reference corpus. *In D. Archer (ed.) What's in Word-list? Investigating Word Frequency and Keyword Extraction*. Oxford: Ashgate. Retrieved from http://www.methodsnetwork.ac.uk/redist/pdf/es1_05 scott.pdf

———. (2016). Introduction to WordSmith Tools; version 7.0. Retrieved from http://lexically.net/downloads/version7/HTML/index.html?getting_started.htm

Sharoff, S. (2014). IntelliText Corpus Queries [Computer Software]. Retrieved from http://corpus.leeds.ac.uk/itweb/htdocs/Query.html

Sinclair, J. (1991). Corpus concordance collocation. Oxford: OUP.

——— (2005). Corpus and text—basic principles. In *Wynne M. (Ed.) Developing linguistic corpora: A guide to good practice*, 1–16.

Sketch Engine. (2015). Retrieved from http://www.sketchengine.co.uk/

Stubbs, M. (1996). Text and corpus analysis. Oxford: Blackwell.

Suchomel, S. & Brandejs, M. (2014). Heterogeneous queries for synoptic and phrasal search. In *CLEF2014 Working Notes*. Sheffield, UK: CEUR, Aachen University. 1017-1020. Retrieved from http://ceur-ws.org/Vol-1180/CLEF 2014wn-Pan-SuchomelEt2014.pdf

———. (1999). Society, education and language: the last 2,000 (and the next 20?) years of language teaching. *Plenary lecture given at the 32nd Annual Meeting of the British Association for Applied*

Linguistics, University of Edinburgh, September 1999.

Swift, M.; Allen, J. & Gildea, D. (2004). Skeletons in the parser: Using a shallow parser to improve deep parsing. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*.

Tanabe, L. & et al. (1999). MedMiner: An Internet Text-Mining Tool for Biomedical Information, with Application to Gene Expression Profiling. *BioTechniques*, 27(6). 1210-1217. Retrieved from http://www.ncbi.nl m.nih.gov/staff/lsmith/papers/tanabe99.pdf

Teubert, W. (2000). A province of a federal superstate, ruled by an unelected bureaucracy: Keywords of the Eurosceptic discourse in Britain. In: *A Musolff, C. Good, P. Points and R. Wittlinger (eds.) Attitudes towards Europe: Language in the unification process.* Aldershot: Ashgate. 45-86.

The Forensics Informatics Biometrics Repository (FIB-R): The open source forensic database. (2015). Retrieved from http://fib-r.com/

The International Corpus of English (ICE). (2015). Retrieved from http://ice-corpora.net/ice/

The Penn Treebank Project. Retrieved from https://www.cis.upenn.edu/~treebank/

The Quranic Arabic Corpus. Retrieved from http://corpus.quran.com/

Thesuarus.com. (2015). Retrieved from http://www.thesaurus.com/

Thompson, C. & Mooney, R. (2003). Acquiring word-meaning mappings for natural language interfaces. *Journal of Artificial Intelligence Research*, 18. 1-44

Thompson, H. & McKelvie, D. (1997). Hyperlink semantics for standoff markup of read-only documents. *Proceedings of SGML Europe*. Retrieved from http://www.ltg.ed.ac.uk/~ht/sgmleu97.html

TREC (Text REtrieval Conference), National Institute of Standards and Technology (NIST). (2015). Retrieved from http://trec.nist.gov/

Van Mol, M. & Paulussen, H. (2001). AraLat: a relational database for the development of bilingual Arabic dictionaries. In S. Lee (Ed.), *Proceedings of Asialex 2001, Asian Bilingualism and the Dictionary* (pp. 206–11). Seoul, August 2001. (https://www.kuleuven-kulak.be/~hpauluss/DOC/asialex.pdf)

Van Mol, M. (1998). Variatie in Modern Standaard Arabisch in radionieuwsbulletins, een synchronisch descriptief onderzoek naar het gebruik van complementaire partikels, (English title: Variation in Modern Standard Arabic in radio news broadcasts, a synchronic descriptive investigation into the use of complementary particles), Ph.D. dissertation, Leuven, 292 p.

———. (2000). Exploring annotated Arabic corpora, preliminary results. In: *Corpora and natural language processing*: *Proceedings of the International Conference on Artificial and Computational Intelligence for Decision, Control, and automation in engineering and industrial applications*.

———. (2000). The development of a new learner's dictionary for Modern Standard Arabic: the linguistic corpus approach. In *U. Heid, S. Evert, E. Lehmann & C. Rohrer (Eds.), Proceedings of the ninth EURALEX International Congress* (pp. 831–36). Stuttgart, 8–12 August. (https://ilt.kuleuven.be/arabic/pdf/A012.pdf)

———. (2003). Variation in Modern Standard Arabic in radio news broadcasts, a synchronic descriptive investigation into the use of complementary particles. Peeters. Belgium.

VArieng: research unit for variation, contacts and change in English. (2014). Retrieved from http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/

Versley, Y. & Panchenko, Y. (2012). Not just bigger: Towards better-quality Web corpora. *Proceedings of the seventh Web as Corpus Workshop (WAC7), Lyon, France*. pp. 45-52. Retrieved from

https://sigwa c.org.uk/raw-attachment/wiki/ WAC7/wac7-proc.pdf

Wilson, J. et al. (2010). Advanced corpus solutions for humanities researchers. *Proceedings of PACLIC 24*, Sendai, Japan.

Winter, T. N. (1999). Roberto Busa, S. J., and the Invention of the Machine-Generated Concordance. Faculty Publications, Classics and Religious Studies Department. 70.

Wynne, M. (1997). Processing large text corpora. In: *A course in the UNIX operating system*. Retrieved from http://users.ox.ac.uk/~martinw/unix/index .html

Xiao, R. (2010). Corpus Creation. In: *Handbook of Natural Language Processing*. London: CRC Press; Taylor & Francis Group.

Xu, J.; Fraser, A. & Weischedel, R. (2002). Empirical Studies in Strategies for Arabic Retrieval. In: *SIGIR 2002, Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. 269-274.

Zaghouani, W. & Dukes, D. (2014). Can crowdsourcing be used for effective annotation of Arabic? *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Retrieved from http://www.l rec-conf.org/proceedings/lrec2014/pdf/43 1_Paper.pdf

Zawaydeh, B. & Saadi, Z. (2006). Orthographic Variations in Arabic Corpora. Retrieved Jun 19, 2013 from http://www.basistech.com/pdf/orthog raphic-variations-in-arabic.pdf

معجم المصطلحات

(A)

Adversative أداة استدراك

Anaphor الإحالة اللغوية القبلية

Anaphoric annotation تحشية العلاقات الإضمارية

Annotated Corpora المدونات اللغوية المَحْشُوَّة

Annotation التحشية

Antecedent الكلمة المذكورة قبلًا

Authentic texts النصوص الواقعية/ الفعلية

(B)

Balance التوازن

Bidirectional corpora المدونات اللغوية ثنائية الاتجاه

Bioinformatics المعلوماتية الحيوية

Bracketing تقويس الكلمات/ وضع الكلمات بين أقواس

(C)

Cataphor الإحالة اللغوية البَعدية

Chi-Square test اختبار مربع كاي

Chunks مقاطع

Clitic words الكلمات ذات الحروف المنفصلة

Cluster analysis تحليل التجمعات العنقودية

Clustering تحليل التجمعات العنقوية

Clusters التجمعات العنقودية

Colligation التلازم المعجمي النحوي

Collocates المتلازمات اللغوية/ المتصاحبات اللغوية

Collocation التلازم اللغوي

Collocational behavior السلوك التلازمي

Combinations of words مجموعات من الكلمات

Comprehensiveness الشمولية

Concordance كشاف النصوص

Concordancers برمجيات تكشيف النصوص

Concordancing تكشيف النصوص

Constituents مكونات الجملة

Content words كلمات المحتوى

Control texts النصوص الضابطة

Coreference annotation تحشية المصاحبة المرجعية

Corpora المدونات اللغوية

Corpus Encoding Standard (CES) معيار تشفير المدونات اللغوية

Corpus Linguistics لسانيات المدونات اللغوية

Corpus of corpora مدونة المدونات اللغوية

Corpus المدونة اللغوية

(D)

Data mining التنقيب عن البيانات

Delimit words الكلمات ذات الحروف المتصلة

Descriptive lexicons المعاجم الوصفية

Diachronic Corpora المدونات اللغوية التعاقبية

Dice Coefficient معامل ارتباط دايس

Disambiguation tagging إزالة اللبس عن الكلام

Discourse analysis تحليل الخطاب

Discourse لغة الخطاب

(E)

Educational Corpora المدونات اللغوية التعليمية

Egyptian Colloquial العامية المصرية

Embedded annotation التحشية المتضمنة

Empty words الكلمات المفرغة

Error tagging وسم الأخطاء

(F)

Forensic Informatics المعلوماتية الجنائية

Functional words الكلمات الوظيفية

Future particles الأدوات الدالة على المستقبل

(G)

General Corpora المدونات اللغوية العامة

Genome الجينوم

Genre النوع الأدبي

Grammatical tagging الوسم القواعدي

Grammatical words الكلمات النحوية

(H)

Hapaxes — الكلمات التي تتكرر مرةً واحدةً فقط

Historic Corpora — المدونات اللغوية التاريخية

Hypertexts — النصوص الفائقة

(I)

Idiomatic expressions — التعبيرات الاصطلاحية

Information extraction — استخراج المعلومات

Invented texts — النصوص الُمخَلَّقة

(K)

Keyness — المفتاحية

Keyword dispersion — تشتت الكلمات المفتاحية

KeyWord In Context (KWIC) — تكشيف الكلمات المفتاحية في السياق

Keywords generation — الكلمات المفتاحية

Knowledge Discovery in Databases (KDD) — الكشف عن المعرفة في قواعد البيانات

(L)

Language variation — التنوع اللغوي

Latent semantic indexing — التكشيف الدلالي الكامن

Learner Corpora — مدونات الدارسين (للغات)

Lemma — الأسرة اللغوية

Lemmatization — تجريد الكلمات/ رد الكلمات إلى أُسَرِها اللغوية

Lexical gap–filling — تعبئة الفجوة المعجمية

Lexical priming — التهيئة المعجمية

Lexical words — الكلمات المعجمية

Lexico–grammatical pattern — التركيب المعجمي النحوي

Lexis and grammar — المفردات والنحو

Light stemming — التجريد الخفيف للكلمات

Likelihood Ration Functions — اختبار دوال الاحتمالات

Loan words — الكلمات الدخيلة

Localization — توطين اللغة

(M)

Machine–readable — مقروءٌ آليًا

Marked-up Corpora المدونات اللغوية المُرَمَّزة

Marking-up الترميز

Meta Corpora ما وراء المدونات اللغوية

Metadata الميتاداتا/ ما وراء البيانات

Modern Standard Arabic (MSA) العربية القياسية الحديثة

Monitor Corpora المدونات اللغوية الراصدة

Monolingual أحادي اللغة

Morphology annotation التحشية الصرفية

Morpho-syntactic tagging الوسم الصرفي النحوي

Mutual Information Coefficient معامل المعلومات المتبادلة

(N)

Nesting التجميع

n-grams تحليل المتتابعات اللفظية

(O)

Ontology الأنطولوجيا

Optical Character Reader (OCR) قارىء المحارف البصرية

(P)

Parallel corpora المدونات اللغوية المتوازية

Parsing التحليل الإعرابي

Part-Of-Speech (POS) tagging وسم أقسام الكلمات

Pattern and meaning التركيب اللغوي والمعنى

Pattern grammar النحو النمطي

Pattern القالب اللغوي

Pedagogical Corpora المدونات اللغوية التربوية/ التدريسية

Phonetic alphabet الألفباء الصوتية

Phrase structure trees أشجار تركيب العبارة

Phraseology الأسلوب العباري للكلمات

Plagiarism انتحال المعلومات العلمية

Plain text format الشكل البسيط للنصوص

Postcedent الكلمة المذكورة بعد ذلك

Pragmatic annotation التحشية البراجماتية أو التداولية

Prefixes السوابق

المعاجم الإرشادية Prescriptive lexicons

التحشية الموجهة نحو المشكلات Problem-oriented annotation

الخصائص العروضية Prosodic features

(R)

الأساس المنطقي Rationale

المدونات اللغوية الخام Raw Corpora

المدونة اللغوية المرجعية Reference corpus

المُحال إليه Referent

السجل اللغوي Register

التمثيل Representativeness

المدونات اللغوية البحثية Research Corpora

التحشية محددة البحث Research specific annotation

الجذر Root

(S)

أخذ العينة Sampling

المسح الضوئي Scanning

التحشية الدلالية Semantic annotation

التحليل الإعرابي الدلالي Semantic parsing

العروض الدلالي Semantic prosody

لغة الترميز المُعَمَّمَة القياسية SGML

علم اللغة الاجتماعي Sociolinguistics

الأفعال الكلامية Speech acts

تحويل الكلام إلى نصوص Speech-to-text

تقنين الهجاء Spelling normalization

التحشية القائمة بذاتها Stand-alone annotation

الجذع Stem

تجريد الكلمات/ رد الكلمات إلى جذوعها Stemming

الكلمات المستثناة Stop words

الأسلوبية Stylistics

اللواحق Suffixes

النصوص المشكوك فيها Suspected texts

التحليل التركيبي Syntactic analysis

المقيدات النحوية Syntactic restrictions

(T)

الوسم Tagging

استخراج الكلمات Term Extraction

مجموعة الاختبار Test collection

المدونات اللغوية الاختبارية Test Corpora

مجموعة الاختبار Test suits /collection

Text Encoding Initiative (TEI) مبادرة تشفير النصوص

المكنز Thesaurus

هيكل الكلمة Token

تجزيء النصوص إلى هياكل من الكلمات Tokenization

نسخ النصوص Transcripting

بنوك اشجار النصوص Treebanks

بتر الكلمات Truncation

الكلمات الفريدة Types

(U)

نظام الشفرة الموحدة Unicode

المدونات اللغوية أحادية الاتجاه Unidirectional corpora

(V)

النوع اللغوي Variety

(W)

معامل الغرابة Weirdness Coefficient

تكشيف الكلمات Word indexing

قائمة الكلمات Word List

وضع الكلمات في قوائم Word listing

وسم معاني الكلمات Word sense tagging

معاني واستخدامات الكلمات Words meanings and uses

البيانات المكتوبة Written data

(X)

لغة الترميز القابلة للتمديد XML

(Z)

قانون زيف Zipf's Law

الكشاف

الأب روبرتو بوسا (رجل دينٍ مسيحيٍ) ٢٩
ابن القيم الجوزية ١٣٢، ١٤٩
الأجرومية – لغةٌ عربيةٌ ١٠٦
الأحاديث النبوية، تجميع نصوص ٣٠
احتمالات سجل الأداء - إحصاءٌ ١٢١، ١٢٤
المكتبات، اختصاصيو ١٥٥
اختيار العينة – مناهج علميةٌ ٣٣، ١٥٤
الأخطاء الشائعة، معاجم ١٠٤
الأخطاء اللغوية، تحليل ١٠٤
الأخطاء اللغوية، وسم ٥٩، ٧٧
آدم كيلجاريف (باحثٌ لغويٌ) ١١٠، ١٤٢
أدوات الاستدراك (كلماتٌ) ١٣٦
أدوات البحث ٢٧، ١٥٤
الأدوات التكميلية (كلماتٌ) ١٣٩، ١٤٠
أدوات الربط (كلماتٌ) ١٣٦، ١٣٧
أدوات معالجة المدونات اللغوية العربية، برنامج ١٢٠، ١٢٣، ١٢٥
أنظر أيضًا: غَوَّاص
خَوَاص
آرالات (قاعدة بياناتٍ علاقيةٍ) ١٤٣
الآريب، دار نشر ١٤٩
إزالة اللبس - معالجةٌ لغويةٌ ٧٠
استخراج المصطلحات ١١٤، ١٥٢
استخراج المعلومات ٤٩، ٦٨
استخراج النصوص ١٣٥
استراتيجيات البحث ١٥٠
الاسترجاع الآلي أحادي اللغة - معالجة نصوصٍ ١٤٩، ١٥٠
الاسترجاع الآلي متعدد اللغات - معالجة نصوصٍ ١٤٩، ١٥٠
استرجاع المعلومات ٢٩، ٤٣، ٤٨، ٥٠، ١٤٨، ١٥٥
استرجاع المعلومات، دراسات ١٣١
استرجاع المعلومات، تجارب ٤٨
الأسر اللغوية - صرفٌ ٦٧، ١١٦
الأسلوب العباري للكلمات - تحليلٌ لغويٌ ٣٩، ٤٠، ٨٦، ١٢٣
الأسلوبية، دراسة ٥٠
الأسلوبية (علمٌ) ٢٠، ٦٨، ٧٦، ٧٧، ٨٦
الإعراب التركيبي - تحليلٌ لغويٌ ٦٧
أقسام الكلمات ١٠٢، ١١٩
أقسام الكلمات، وسم ٦٧، ٧٠-٧٥
آكونكورد، برنامج ١٢٦

الألفباء الصوتية ٥٥
إم بي ثري - أشكال حفظ النصوص المنطوقة ١٠٥
الانتحال العلمي ٥٠، ٥٣
أنظر أيضًا: السرقات العلمية
الإنترنت ٩٩، ١٠٥، ١٣٤، ١٣٨، ١٤١، ١٥١
الإنترنت، استثمار ١٧
آنتكونك، برنامج ١٢٠، ١٢٦
الأنطولوجيا (تحليلٌ منطقيٌ للأشياء) ١٠٨، ١٥٢
الأنطولوجيا الدلالية ١٠٨
الأنواع الأدبية ٣٤، ٤١، ٩١–٩٣، ١٠٣، ١١٦، ١١٧، ١٣٢، ١٣٣
الأنواع اللغوية ٩٢، ١٣٢
الأهرام (صحيفةٌ مصريةٌ) ١٠٣، ١٣١، ١٤٠
ائتلاف البيانات اللغوية (منظمةٌ) ١٤٩
الآيزو (مواصفةٌ قياسيةٌ) ٥٧
بتراء (صحيفةٌ أردنيةٌ) ١٣١
البطاقات المثقبة ٢٩
بَكوالتر، برنامج ٧١، ٧٢، ١٠٢، ١٥٠
بنك الأشجار النحوي - معالجةٌ لغويةٌ ١٠٧
بنوك أشجار النصوص - معالجةٌ لغويةٌ ٦٨، ٧٣، ٧٤
بي دي إف - حفظ نصوصٍ ١٠٤، ١٠٥، ١٣١
البيانات، معالجة ٦٣، ٦٥، ٦٦، ٧٩
التباين - ظواهر لغويةٌ ١٣٩، ١٤٠، ١٤٢
تجارب كرانفيلد - استرجاع معلوماتٍ ٤٨
التجريد الخفيف للكلمات - معالجة نصوصٍ ١٥٢
تجريد الكلمات - معالجةٌ لغويةٌ ٦٧، ٧٣، ١٤٨، ١٥٠
تجزيء النصوص إلى هياكل من الكلمات - معالجةٌ لغويةٌ ٧٢
التجمعات العنقودية، تحليل ٢٠، ٦٣، ٦٥، ٨٤، ٨٦
التجميع - معالجة نصوصٍ ١٠١، ١٣٨
التحشية الأسلوبية ٦٨، ٧٦، ٧٧
التحشية الآلية ٦٩، ٧٠
التحشية البراجماتية ٦٨، ٧٦
أنظر أيضًا: التحشية التداولية
التحشية الخطابية ٦٨
التحشية الدلالية ٦٨، ٧٥
التحشية الصرفية ٦٧، ١٠٦

التحشية الصرفية النحوية ١٠٦
التحشية الصوتية ٦٧
التحشية العَروضية ٦٧
التحشية القائمة بذاتها ٧٨
تحشية المدونات اللغوية ١٩، ٦٣، ٦٥-٦٧، ٦٩، ٧٠
تحشية المدونات اللغوية، أنواع ٧٠
تحشية المدونات اللغوية، طرق ٧٠
تحشية المصاحبة المرجعية ٦٨، ٧٥
التحشية الموجهة نحو المشكلات ٧٨
التحشية النحوية ١٠٦
التحشية اليدوية ٧٠
التحشية شبه الآلية ٧٠
التحشية - معالجةٌ لغويةٌ ٢٥، ٦٧، ٦٨، ٧٨، ١٠٦
التحقق من النصوص ٥٦، ٧٧
التحليل الإعرابي ٦٨، ٧٣-٧٥
تدريس وتعلم اللغات ٣٩، ١٣٦
الترجمة ٤٦، ٤٧، ٨٠، ١٤٥، ١٤٦، ١٥٠، ١٥٢
الترجمة الآلية ١٣١
الترجمة الآلية، متطلبات ١٤٥
تردد الكلمات - معالجةٌ لغويةٌ ٢٩، ٤٣، ٤٤، ٨٢-٨٤، ١٤٧
ترِك، مؤتمر ١٤٩
ترِك، مجموعات نصوص ٤٨
التركيب المعجمي النحوي ١٣٦
التركيبة المعجمية ٤٧
الترميز - معالجةٌ لغويةٌ ٢٥، ٦٣، ٦٥-٦٧، ٧٨، ١٠٠، ١٣٥
التشتت، معيار ١٢٢
تشرين (صحيفةٌ سوريةٌ) ١٣١
تشفير النصوص ٥٦
تشومسكي (عالم لغةٍ) ٨٦
تعبئة الفجوة المعجمية ١٣٨
التعبيرات الاصطلاحية ٣٣، ٤٠، ٤٣، ٨٦
التغطية، معيار ٤٨
التقنين ٧٩، ١٥٤
تقنين الهجاء ١٥٠
تقويس الكلمات - معالجةٌ لغويةٌ ٦٨
التكشيف الآلي ١٤٨، ١٤٩
التكشيف الدلالي الكامن ٤٩
تكشيف الكلمات المفتاحية ٦٥، ٧٩، ٨٢، ٩١، ١٥٥
تكشيف الكلمات المفتاحية في السياق ٢٠، ٦٣، ٧٩، ٨١
تكشيف اللغة الطبيعية ٧٩

تكشيف المدونات اللغوية، مشكلات فنية ٦٣، ٨٢
تكشيف المدونات اللغوية، مشكلات لغوية ٦٣، ٨٢
تكشيف النصوص ٢٨، ٣٠، ٣١، ٧٩-٨٢، ١١٣
تكشيف الوثائق ١٥٣
التكشيف اليدوي ١٤٨، ١٤٩
التلازم اللغوي، تحليل ٧٠، ١٢٢
التلازم اللغوي - ظواهرة لغويةٌ ٤٠، ٤٣، ٩٩، ١٠٠، ١٣٨
أنظر أيضًا: التصاحب اللفظي
المصاحبة اللفظية
التلازم اللفظي
تمثيل المعرفة ١٠٨
التمثيل، معيار ٣٢، ٣٣، ٥٣، ١٠٠، ١٠١، ١٥٤
التمثيلات الخاصة - نصوص ٥٦
التنقيب عن البيانات ٤٩
التنوع اللغوي ٤١
التهيئة المعجمية ١٣٨
التوازن، معيار ٣٣، ٤٥، ١٠١
التوجهات الفكرية، تحليل ١٤٦
التوجهات الفكرية، دراسة ٣٥، ٤٧
التوجهات الفكرية (فِكرٌ) ٢٠، ٤٧، ١٤٦
التوزيع التكراري - إحصاءٌ ١٣٠، ١٣٢
توطين اللغة ١٤٠
جامعة أكسفورد ٨١، ١٢٠
جامعة الإمام محمد بن سعود الإسلامية ١٠٤
جامعة الملك فيصل ٩٣
جامعة إلينوي بإربانا شامبين ١٥٥
جامعة أم القرى ٩٣
جامعة برجهام يَنج الأمريكية ٢٧، ١٠٣، ١١٦، ١١٩
جامعة فيكتوريا ويلنجتون بنيوزيلاندا ٢٦
جامعة ليدز ٦٠، ٧٤، ١٠٩، ١٣١
جامعة ليفربول ١٢٠
جامعة نورث كارولاينا بشابل هيل ١٥٤
الجذور - صرفٌ ١٠٠، ١٠٢، ١١٩، ١٤٣، ١٤٨-١٥٠
الجذوع - صرفٌ ٧٣، ٩٦، ١٠٠، ١٠٢، ١٤٨، ١٤٩
الجزيرة (صحيفةٌ سعوديةٌ) ١٣١
الجزيرة (قناةٌ تليفزيونية) ١٥٠، ١٥١
جلوسب (معجمٌ إلكترونيٌ متعدد اللغات) ٤٧

جمعية المكتبات السلوفينية ١٥٤
حقوق الملكية الفكرية ٩٤
الحياة (صحيفةٌ لبنانيةٌ) ١٠٣، ١٣٠، ١٣١
الخطاب، أنماط ١٤٧
الخطاب، تحليل ٤٧
دار نشر جامعة أوكسفورد ٤٤
دار نشر جامعة كمبريدج ٤٤
دبلن كور (إطارٌ معياريٌ) ٦٧
الدلالة– علم لغةٍ ٤٢، ١٢٤، ١٣٦
ديلوورث باركينسون (باحثٌ لغويٌ) ١٠٣، ١٤٠
الرأي العام (صحيفةٌ كويتيةٌ) ١٣١
راية القطرية (صحيفةٌ) ١٣١
روز اليوسف (صحيفةٌ مصريةٌ) ٩٣
زايرا، برنامج ٦٩
الزحف – تكشيف وتجميع محتوى ١٣٥
سارا، برنامج ٦٩، ٧٢
الساق – صرفٌ ١٠٢، ١٥٤
السجل اللغوي – علم لغةٍ اجتماعيٍ ٤١
السحب الضوئي – قراءةٌ آليةٌ للحروف ١٣٣
سفير (صحيفةٌ لبنانيةٌ) ١٣١
سكتش إنجين (مدونةٌ لغويةٌ) ٢٧، ٤٤، ٥٢، ١١٠–١١٥، ١٢٠
السلوك التلازمي – علم لغةٍ ١١٢، ١٣٦، ١٣٧
السوابق – صرفٌ ٦٧، ٩٦، ١٠٢، ١٥٠
الشبكة العنكبوتية ١٨، ٢٦، ٥٥، ٩١، ٩٣، ١٠٠، ١٠١، ١١٠، ١١٧، ١٣١، ١٣٤، ١٣٥، ١٥٢
الشروق (صحيفةٌ مصريةٌ) ١٠٣، ١٣٦
الشمولية، معيار ٣٣، ٤٨
صامويل جونسون (باحثٌ لغويٌ) ٤٣
صيد الفوائد (مكتبةٌ رقميةٌ) ٩٣
عبد الله الفيفي (باحثٌ لغويٌ) ٢٦، ١٠٤
عبد المحسن الثبيتي (باحثٌ لغويٌ) ٨١، ٨٤، ١٥١–١٥٣
عربي كوربص، مدونة ٢٧، ١٠٣، ١٠٤، ١٣٦
العربية القياسية الحديثة ٩٣، ١٣٢، ١٤١
العربية القياسية الحديثة - خصائص معجميةٌ ١٤٠
العروض الدلالي ١٣٦، ١٣٧
العروض الدلالي – علم لغة ١٣٦
علم اللغة ٢٨، ٣٨، ٧٦، ٨٠، ١٥٣
علم اللغة الاجتماعي ٤٢، ١٣٩، ١٤٠

علم اللغة التطبيقي ٢٠، ٣٨، ١٠٤، ١٣٦، ١٤٠
علم المعلومات ٢٨، ٢٩، ٤٨، ٤٩، ٨٠، ١٤٨، ١٥٤
الغد (صحيفةٌ أردنيةٌ) ١٠٣
غرف الدردشة الإلكترونية – مصادر لغةٍ ١٣٣
الفصحى التراثية ٩٣
فنسنك (مستشرقٌ هولنديٌ) ٣١
قارئ المحارف البصرية، برنامج ٥٤
قانون زيف – إحصاءٌ ٤٤، ١٣٠– ١٣٢
القانون، مجال ٥٢
القديس توما الأكويني – مصادر لغةٍ ٢٩
القرآن الكريم ٣٠، ٣١، ٥٨، ١٠٦، ١٠٨، ١٣٣، ١٤١، ١٥٢
القرآن الكريم – مصادر لغةٍ ٣٠
القرآن الكريم – معالجةٌ لغويةٌ ١٠٦
قواعد بيانات النصوص ٢٣، ٢٨، ٢٩
قوائم الاستثناء – معالجةٌ لغويةٌ ١٢٤
قوائم الاعتبار – معالجةٌ لغويةٌ ١٢٤
قوائم تردد الكلمات – معالجةٌ لغويةٌ ٢٩، ٨٢، ٨٣، ١٠٠
كاكاو (إطارٌ معياريٌ) ٦٧
كتاب نجوم الفرقان ٣١
كتب جوجل (مكتبةٌ رقميةٌ) ١١٩
الكشاف السياقي ٩٥، ٩٦، ٩٨، ٩٩
كشاف الكلمات المفتاحية في السياق ١٢٢
كشاف النصوص ١١٣
كشاف نصوص أشعار ماثيو أرنولد ٣٠
كشاف نصوص أعمال وليم شكسبير ٣٠
الكشف عن المعرفة في قواعد البيانات ٤٩
الكلمات الدخيلة – ظواهر لغويةٌ ٤٢، ١٤٢
الكلمات الفريدة ١٢٥
كلمات المحتوى – معالجةٌ لغويةٌ ٤٠
الكلمات المستثناة ٨٠، ١٠٠
الكلمات المعجمية ٤٠
الكلمات المفتاحية ٢٩، ٤٧، ٦٥، ٨١-٨٤، ٩١، ١١٨، ١١٩، ١٢٢، ١٢٥، ١٥٣- ١٥٥
الكلمات المفتاحية، تعريف ٨١
الكلمات المفرغة ٤٠
الكلمات المقترَضة – ظواهر لغويةٌ ٤٢
الكلمات النحوية ٤٠
الكلمات الوظيفية ٤٠، ٨٣
الكلمات المفتاحية، معالجة ٤٨، ٨٣، ٨٤، ١٢٣
اللانسوية – ظواهر اجتماعيةٌ ١٤٧، ١٤٨

اللسانيات، علم ١٣٧
لسانيات المدونات اللغوية، علم ١٣٧
لغة الترميز القابلة للتمديد (برمجةً) ٦٧، ٧٢، ١٠٤، ١٠٨
أنظر أيضًا: إكس إم إل
لغة الترميز القابلة للتمديد، استخدام ١٠٠
لغة الترميز المُعَمَّمَة القياسية (برمجةٌ) ٦٧، ٧٢، ٧٨، ١٣٥
أنظر أيضًا: إس جي إم إل
لغة الخطاب ٤١، ١٣٧
اللغة الطبيعية ٤٧، ٦٠، ١٣٥، ١٤٤، ١٥٠
اللغة الطبيعية، معالجة ١٤٤
اللغة العربية ١٨، ١٩، ٣٠، ٣٩، ٤٢، ٤٦، ٤٧، ٦٨، ٧١، ٧٢، ٧٦، ٧٧، ٧٩، ١٠١، ١٠٤، ١٠٦، ١١٠، ١٣١، ١٣٢، ١٤٢، ١٤٣، ١٥٠، ١٥٥
اللغة العربية المعاصرة ١٠١
اللغة العربية، استعمال ١٣٩
اللغة العربية، تدريس ١٣٧
اللغة العربية، طلاب ١٠٤
اللغة العربية، نظم استرجاع ١٤٩
لغة تهيئة النصوص الفائقة (برمجةٌ) ١٣٠، ١٣٤
أنظر أيضًا: إتش تي إم إل
اللواحق - صرفٌ ٦٧، ٩٦، ١٠٢، ١٥٠
ما وراء المدونات اللغوية ٦٠
مارك ديفيس (باحثٌ لغويٌّ) ١١٦، ١١٩
مارك فان مول (باحثٌ لغويٌّ) ١٣٩، ١٤٢
مايك سكوت (باحثٌ لغويٌّ) ٨١، ١٢٠
مايكرو-إيرس، نظام استرجاع ١٤٨
مبادرة تشفير النصوص - معايير ٦٧
المتتابعات اللفظية، تحليل ١١٥، ١٢٤، ١٣٠
المتلازمات اللغوية ٣٢، ٨٦، ١٠٠، ١١١، ١١٥، ١١٦، ١٢٢، ١٢٥، ١٤٥، ١٥١، ١٥٢
المتلازمات اللغوية، تحليل ١٤٧
مجموعة اختبار (أداةٌ منهجيةٌ) ٤٨
مجموعة الاهتمام الخاص باسترجاع المعلومات، مؤتمر ١٥٠
المحدد الموحد لمكان المصدر ٩٢
أنظر أيضًا: يو آر إل
محركات البحث ٩٥

مخطط الكلمات - معالجةٌ لغويةٌ ١١١، ١١٢
المداخل الكشفية ٧٩، ١٤٨، ١٤٩
مدونات الدارسين ٥٩
المدونات اللغوية ١٨، ١٩، ٢٦-٢٩، ٣١-٣٤، ٣٨-٤٧، ٤٩-٥٣، ٥٥-٦٠، ٦٥-٧٦، ٧٨، ٧٩، ٨١، ٨٢، ٩١، ٩٢، ١١٠، ١١٢، ١١٤-١١٨، ١٢٢، ١٢٣، ١٢٦، ١٣٠، ١٣٢، ١٣٣، ١٣٦، ١٣٩، ١٤٠، ١٤٢-١٤٤، ١٤٦، ١٤٨، ١٥١-١٥٥
المدونات اللغوية أحادية اللغة ٤٦
المدونات اللغوية الأحادية ٥٧
المدونات اللغوية الاختبارية ٥٧
المدونات اللغوية البحثية ٥٧
المدونات اللغوية التاريخية ٥٨
المدونات اللغوية التدريسية ٦٠
المدونات اللغوية التربوية ٦٠
المدونات اللغوية التعاقبية ٥٨
المدونات اللغوية التعليمية ٥٩
المدونات اللغوية الخام ٥٧، ٦٧، ٦٩
المدونات اللغوية الراصدة ٥٨
المدونات اللغوية العامة ٥٧
المدونات اللغوية العربية واسترجاع المعلومات، دراسات استخدام ١٤٨
المدونات اللغوية العربية، تقييم ١٣٢
المدونات اللغوية العربية، دراسات استخدام ١٣٦
المدونات اللغوية العربية، معالجة ١٢٠
المدونات اللغوية القائمة على الشبكة العنكبوتية ٢٧، ٩١
المدونات اللغوية المتخصصة ٥٨
المدونات اللغوية المتوازية ٤٦
المدونات اللغوية المتوازية ٥٧
المدونات اللغوية المَحْشُوَّة ٥٧، ٦٩
المدونات اللغوية المرجعية ١٢٥
المدونات اللغوية المرجعية ١٢٦
المدونات اللغوية المرَمَّزة ٥٧، ٦٧
المدونات اللغوية المقارِنة ٥٧
المدونات اللغوية المقارنِة ٤٦
المدونات اللغوية، برمجيات معالجة ١٢١، ١٢٤
المدونات اللغوية، تعريف ٢٥
المدونات اللغوية ثنائية اللغة، معالجة ١٤٥
المدونات اللغوية، خصائص منهجية ٦٠

المدونات اللغوية، دراسات إفادة ١٣٦، ١٣٨، ١٤٢، ١٤٥، ١٤٧-١٤٩، ١٥١، ١٥٣
المدونات اللغوية والترجمة ١٤٥
المدونات اللغوية في تحليل التوجهات الفكرية، دراسات استخدام ١٤٦
المدونات اللغوية، استثمار ٢٠، ٤٠، ٤١، ٤٧، ١٤٥، ١٤٨، ١٥٣
المدونات اللغوية، استخدام منهجي ٦٠
المدونات اللغوية، إفادة ٢٠، ٣٥، ٣٨، ٣٩، ٤٢، ٤٦-٤٨، ٥٠، ٥٢، ١٣٩، ١٥٤
المدونات اللغوية، مهام ٢٥
المدونات اللغوية، مصطلح ومفهوم ٢٧
المدونات اللغوية، إنشاء وإتاحة ١٣٣
المدونات اللغوية، أنواع ٣٥، ٥٧
المدونات اللغوية، بناء وتنقيب ١٣١
المدونات اللغوية، صعوبات ٣٤
المدونات اللغوية، تحليل ٧٠
المدونات اللغوية، ترميز ٦٣، ٦٥-٦٧
المدونات اللغوية، تصميم ١٥٤
المدونات اللغوية، تطبيقات ٣٨
المدونات اللغوية، تكشيف ٨٠
المدونات اللغوية، دراسات استخدام ١٣٦، ١٣٩، ١٤٦، ١٥١، ١٥٣
المدونات اللغوية، دراسات إنشاء وإتاحة ١٣٠، ١٣١، ١٣٣، ١٣٥
المدونات اللغوية، فوائد ٣٣، ٤٣
المدونات اللغوية، متطلبات ٤٤
المدونات اللغوية، مجالات إفادة ٣٥، ٣٨
المدونات اللغوية، مراحل إنشاء ٣٥، ٥٣، ١٣٠
المدونات اللغوية، مميزات ٢٣، ٣١
المدونات اللغوية، مواصفات ٢٣، ٣٢
المدونات اللغوية، نماذج ٩٢، ١٠٠، ١٠١، ١٠٤، ١٠٧، ١٠٨-١١٠، ١١٦، ١١٩، ١٢٠
مدونة استقصاء المدونات اللغوية العربية ١١٠
المدونة الإنجليزية الدولية الراصدة ٥٩
المدونة العربية القرآنية ١٠٧-١٠٩
المدونة القومية العربية ١٣٤
المدونة اللغوية التاريخية للجامعة الأردنية ١٠٠
المدونة اللغوية الدولية لدارسي الإنجليزية ٥٩
المدونة اللغوية الدولية للغة الإنجليزية ٤٥، ٤٦

المدونة اللغوية العربية الدولية لمكتبة الإسكندرية ١٠١، ١٠٣
المدونة اللغوية العربية لعلوم الحاسب ١١٠
المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية ٩٢، ٩٥، ٩٦، ٩٨، ١٢٤، ١٢٦، ١٤٧
أنظر أيضًا: غَوَّاص، برنامج خَوَاص، برنامج
المدونة اللغوية العربية لوكالة الأنباء الفرنسية ١٥٠
المدونة اللغوية القانونية العربية ١٠٩
المدونة اللغوية الوطنية الأمريكية ٣٠
المدونة اللغوية الوطنية البريطانية ٣٠، ٥٧، ١١٧
المدونة اللغوية لجريدة الحياة ١٠٩
المدونة اللغوية لكتب جوجل ١١٧، ١١٩
المدونة اللغوية للأخبار على الشبكة العنكبوتية ١١٨
المدونة اللغوية للإنترنت ١٠٩
المدونة اللغوية للإنجليزية الأمريكية المعاصرة ١١٦
المدونة اللغوية للعربية المعاصرة ١٠٩، ١٣١
المدونة اللغوية لمتعلمي اللغة العربية ٢٠، ٢٦، ٦٠، ١٠٤، ١٠٥
المدونة اللغوية لويكيبيديا ١٠٩
المدونة الممثِلة للسجلات اللغوية الإنجليزية التاريخية ٥٨
مدونة أوساك ١٣٨، ١٥٢
مدونة إنتلتيكست ١٢٠
مدونة براون ٢٩
مدونة دارسي كامبريدج ٦٠، ٧٧
مدونة زاد ١٤٩، ١٥٠
مدونة سينيكا ١٥٣
مدونة قرآني ١٠٨
مدونة كتب جوجل ١١٩
مدونة كوكا ١١٦، ١١٧، ١١٨
مدونة لوفان لمقالات الإنجليزية الأصلية ٥٩
مدونة لونجمان للغة الأمريكية المكتوبة ٥٩
مدونة هيلسينكي للغة الإنجليزية ٥٨
مدونة هيلسينكي للنصوص الإنجليزية ٤٥
مدونة ويلنجتون للإنجليزية النيوزيلاندية المكتوبة ٢٦

مدونة ويلنجتون للإنجليزية النيوزيلاندية المنطوقة ٢٦
مدينة الملك عبد العزيز للعلوم والتقنية ٩٣، ١٢٤
مربع كاي (معاملٌ إحصائيٌ) ٨٣، ١٢٢، ١٢٤
مستودع المعلوماتية الجنائية للقياسات الحيوية ٥٣
المسح الضوئي (قراءةٌ آليةٌ للحروف) ٥٤
مشروع تقويم تحشية الجينوم ٥٢
مصادر المعلومات ١٨، ٩١-٩٣، ١٣٤
مصادر المعلومات، أنواع ٥٨
المصري اليوم (صحيفةٌ مصريةٌ) ١٠٣
المصطلحات الكشفية ٩١، ١٥٠
المعاجم الإرشادية ٤٤
المعاجم الطلابية، تأليف ١٠٤
المعاجم اللغوية، صناعة ٤٢، ٤٤، ٤٥، ٨٠، ٨٦، ١٠٤، ١٤٠-١٤٤
المعاجم الوصفية ٤٤
معالج استفسارات المدونات اللغوية، برنامج ١٢٠
معالجة اللغة الطبيعية، تطبيقات ٩٢
معامل ارتباط الغرابة - إحصاءٌ ١٢٤، ١٢٥، ١٥١
معامل ارتباط دايس - إحصاءٌ ١٢٤، ١٢٥
معامل الارتباط - إحصاءٌ ٩٩
معامل المعلومات المتبادلة - إحصاءٌ ١٢٤
المعجم المفهرس لألفاظ القرآن الكريم ٣١، ٥٨
المعلوماتية الجنائية ٥٢، ٥٣
المعلوماتية الحيوية ٥٢، ١٥٤
معيار تشفير المدونات اللغوية ٦٧
المغرب اليوم (صحيفةٌ) ١٣١
المفاهيم القرآنية، أنطولوجيا ١٠٨
مقالات الأخبار ٤٥
المكانز، إنشاء وإتاحة ٥١
المكانز، صناعة ٥٠، ١٥١، ١٥٥
المكتبات الرقمية ١٥٣
المكتبات، علم ٥٠، ٨٦، ١٥٣-١٥٥
المكتبة الشاملة (مكتبةٌ رقميةٌ) ٩٣
مكتبة هاثي تراست الرقمية ١٥٥
المكنز (أداة بحثٍ) ٥٢، ١١٢
ملفات نوت باد - أشكال حفظ نصوصٍ ١٠٥
مناهج بي سي يو - معالجةٌ لغويةٌ ١٤٧
موقع اتحاد الكتاب العرب ٩٣
مونو كونك، برنامج ٦٩

الميتاداتا ٢٥، ٥٠، ٦٦، ٩٤، ١٠٠، ١٠٥، ١٣٥
النحو العربي، قواعد ١٣٩
النحو، قواعد ١٨، ٢٦، ٤٢، ٥١، ٦٨، ٧١-٧٣، ٧٨، ١٣٦، ١٣٨، ١٤٠، ١٥٥
نسخ النصوص – طرق جمعٍ ٥٥
النسوية – ظواهر اجتماعيةٌ ١٤٧، ١٤٨
نصوص المدونات اللغوية، معالجة ٤٧
النصوص المكتوبة ٢٧، ٢٨، ٣٣، ٥٦، ٦٦، ٦٧، ٧٢، ٨٤، ٩٣، ١٠٥، ١٣٤، ١٤٢
النصوص المنطوقة ٢٨، ٥٥، ٦٦، ٦٧، ١١٦
النصوص، تحليل ١٢٠، ١٤٤
النصوص – طرق حفظٍ ٥٦، ٦٦، ١٠٢
نظم استرجاع المعلومات ٤٩، ٨٦، ١٥٤
أنظر ايضًا: النظم الآلية للاسترجاع
نظم استرجاع المعلومات، اختبار وتقييم ٤٩
نقحرة الكلمات ١٤١
نيفا، برنامج ١٥٥
هياكل الكلمات ٣٢، ٧٢، ١٠١، ١٢٤، ١٣٤
الوثائق العربية ١٥٠
الوثائق العربية، تكشيف واسترجاع ١٤٨
ورشة العمل الدولية الأولى للمكتبات الرقمية في علم الموسيقى ١٥٥
الوزن الصرفي ١٠٢
الوزن (معيارٌ لغويٌ) ١٠٠، ١٥٢
الوسم الصرفي النحوي ٧٠
الوسم القواعدي ٧٠
وضع الكلمات في أسرٍ لغويةٍ ٦٧
الوطن (صحيفةٌ سعوديةٌ) ٩٣
الوطن (صحيفةٌ عمانيةٌ) ١٣١
وكالة الأنباء الجزائرية ١٣١
وكالة الأنباء السعودية ٩٣
وكالة الأنباء العراقية ١٣١
وورد سميث، برنامج ٢٠، ٨٠، ١١٨، ١٢١، ١٢٢، ١٣٣، ١٣٤، ١٣٨

دراسات ١٢

المدونات اللغوية
ودورها في معالجة النصوص العربية

يعمل مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية على تعزيز خدماته في المجالات المتنوعة لخدمة اللغة العربية وعلومها، إذ ينطلق من رؤية موحّدة في أعماله عامة - ومنها برنامج النشر - وذلك بأن يطلق برامجه ودراساته في المجالات التي تفتقر إلى جهود نوعية، أو التي تحتاج إلى تكثيف العمل فيها.

ويجتهد المركز في انتقاء الكتب التي تصدر ضمن هذه السلسلة، بأن تكون مضيفة إلى حقلها المعرفي، ومفتاحا للمشروعات العلمية والعملية، ومحققة لتراكم معرفيٍّ مثرٍ.

وإذ تشيد الأمانة العامة في المركز بجهد مؤلف الكتاب، تأليفاً، وتصحيحاً لمسوداته، ومراجعةً للطباعة، فإنها تدعو الباحثين كافة من أنحاء العالم إلى المساهمة في هذه السلسة، لتتكامل مع سلاسل المركز العلمية الأخرى.

ويسعد المركز بالعمل مع المؤسسات والأفراد المختصين والمهتمين في خدمة لغتنا العربية، وتكثيف الجهود والتكامل نحو تمكين لغتنا، وتحقيق وجودها السامي في مجالات الحياة.



ص.ب ١٢٥٠٠ الرياض ١١٤٧٣
هاتف: ٠٠٩٦٦١١٢٥٨٧٢٦٨ - ٠٠٩٦٦١١٢٥٨١٠٨٢
البريد الإليكتروني: nashr@kaica.org.sa




9 786038 221099